Attentive Captioning without Attention

Kate Saenko

BOSTON UNIVERSITY

UC Berkeley CS 294-131, Special Topics in Deep Learning, March 2017

Problem: Captioning images or video

Image Description

Input image

Output: A close up of a hot dog on a bun.

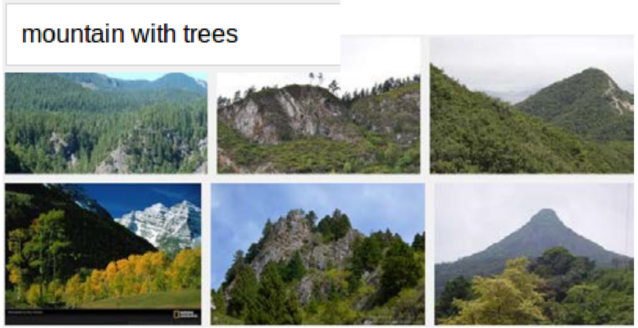
Video Description

Input video

Output: A woman shredding chicken in a kitchen

Applications

Image and video retrieval by content.



Video description service.

Human Robot Interaction

Video surveillance

Today

ICCV15 – end-to-end video captioning

ACM MM16 – multimodal video captioning

CVPR17 – caption-guided video saliency

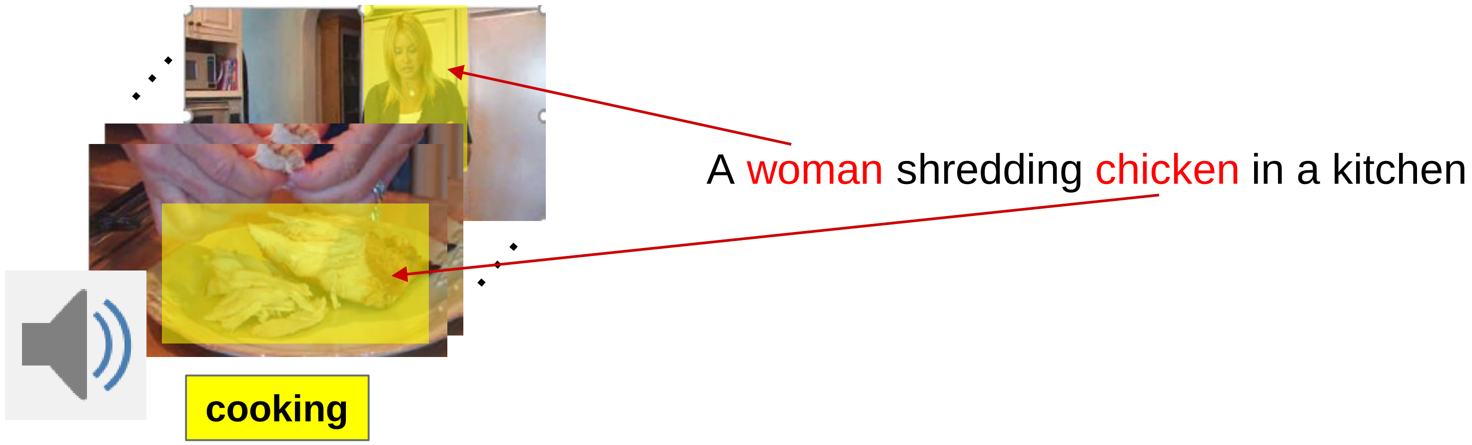


Image Captioning, B.D. (before deep learning)

Language: Increasingly focused on **grounding** meaning in perception.
Vision: Exploit linguistic ontologies to **“tell a story”** from images.

[Farhadi et. al. ECCV’10]

(animal, stand, ground)

[Kulkarni et. al. CVPR’11]

There are one cow and one sky.
The golden cow is by the blue sky.

Many early works on Image Description
Farhadi et. al. ECCV’10, Kulkarni et. al. CVPR’11, Mitchell et. al. EACL’12, Kuznetsova et. al. ACL’12 & ACL’13

Identify objects and attributes, and combine with linguistic knowledge to “tell a story”.

Dramatic increase in interest 2015
(8 papers in CVPR’15)

Video Captioning, B.D. (before deep learning)

[Krishnamurthy, et al. AAAI'13]

[Yu and Siskind, ACL'13]

[Rohrbach et. al. ICCV'13]

- Extract object and action descriptors.
- Learn object, action, scene classifiers.
- Use language to bias visual interpretation.
- Estimate most likely agents and actions.
- Template to generate sentence.

Others: Guadarrama ICCV'13, Thomason COLING'14

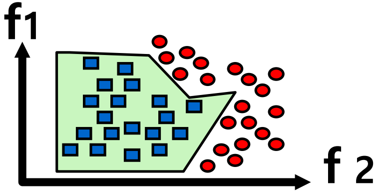
Limitations:

- Narrow Domains
- Small Grammars
- Template based sentences
- Several features and classifiers

Deep Learning Revolution

Input

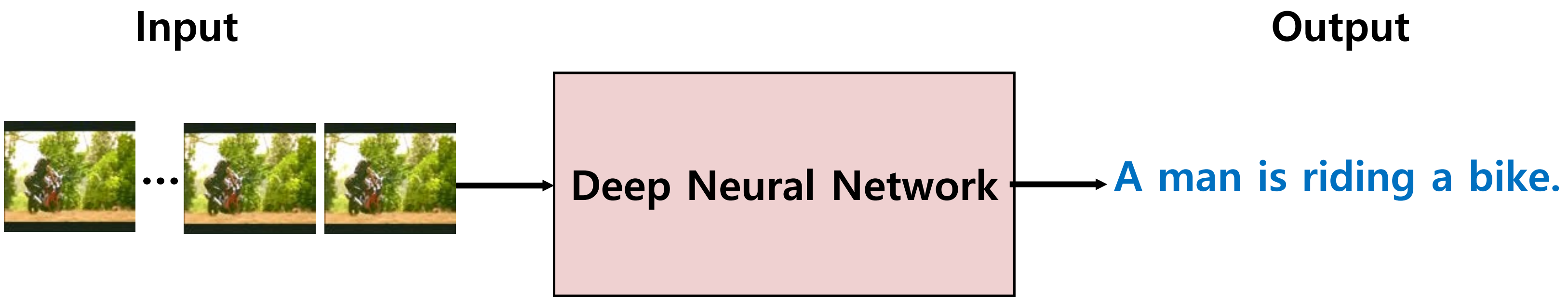
Manual Feature Design → $f_1 \ldots f_n$ → Machine Learning → output



Input

Deep Neural Network → output

Video description: Sequence-to-sequence problem



Deep End-to-End Neural Models based on Recurrent Nets

Deep Neural Network → Sentence

[Donahue et al. CVPR'15] (our work)
[Vinyals et al. CVPR'15]

Deep Neural Network → Sentence

[Venugopalan et. al. NAACL'15]
[Venugopalan et. al. ICCV'15] (our work)

Today

ICCV15 – end-to-end video captioning

ACM MM16 – multimodal video captioning

CVPR17 – caption-guided video saliency

End-to-End Neural Video Description

Subhashini Venugopalan

UT Austin

Jeff Donahue

UC Berkeley

Marcus Rohrbach

UC Berkeley

Raymond Mooney

UT Austin

Trevor Darrell

UC Berkeley

[Venugopalan et. al. NAACL’15] [Venugopalan et. al. ICCV’15]

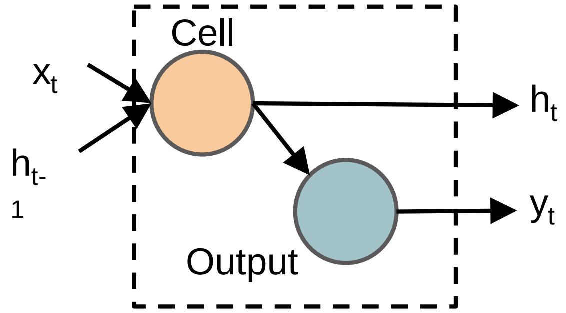
[Background] Recurrent Neural Networks

Successful in translation, speech.

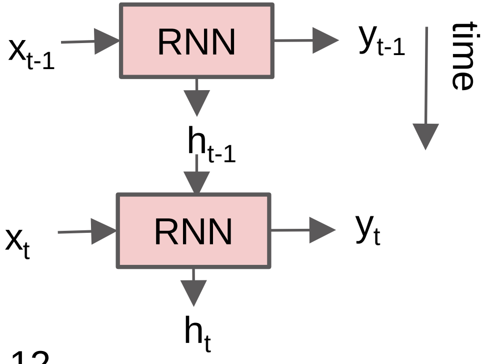
RNNs can map an input to an output sequence.

$$\Pr(\text{out } y_t \mid \text{input}, \text{out } y_0 \ldots y_{t-1})$$

Insight: Each time step has a layer with the same weights.



RNN Unit



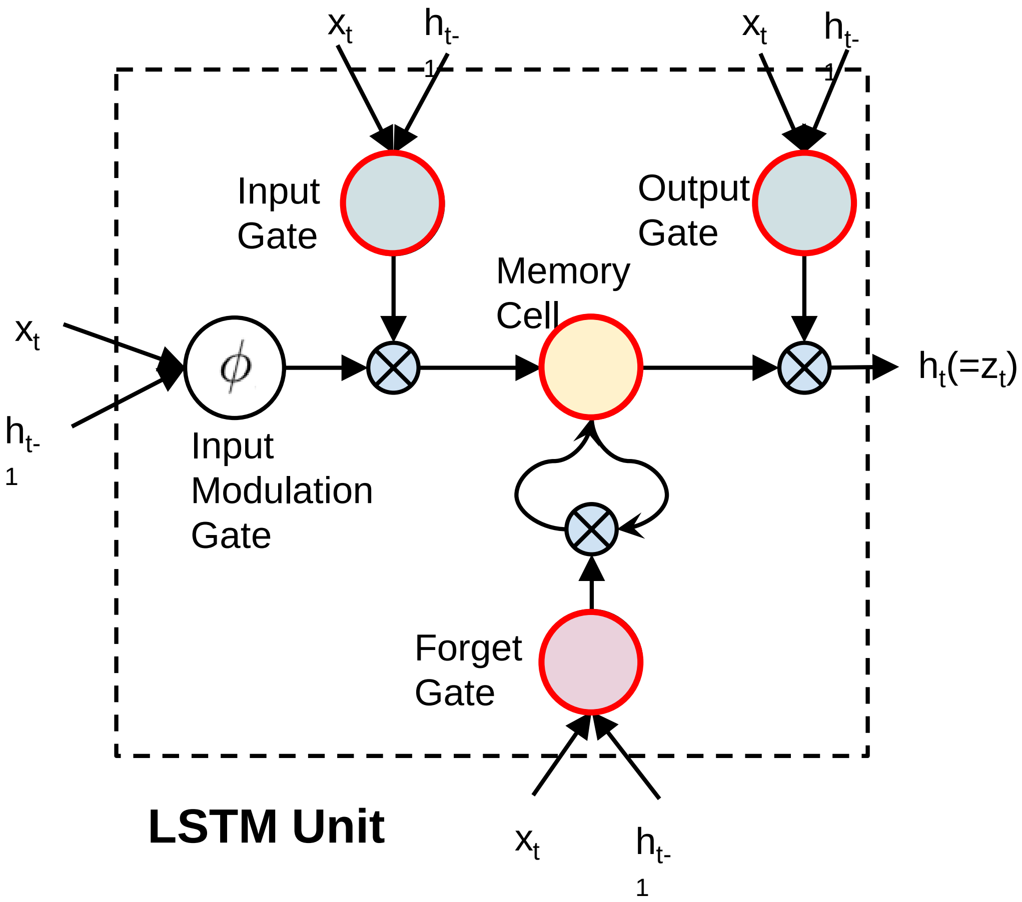
Problems:

1. Hard to capture long term dependencies
2. Vanishing gradients (shrink through many layers)

Solution: Long Short Term Memory (LSTM) unit

[Background] LSTM

[Hochreiter and Schmidhuber ‘97]
[Graves ‘13]



LSTM Unit

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1})$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1})$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{xc}x_t + W_{hc}h_{t-1})$$

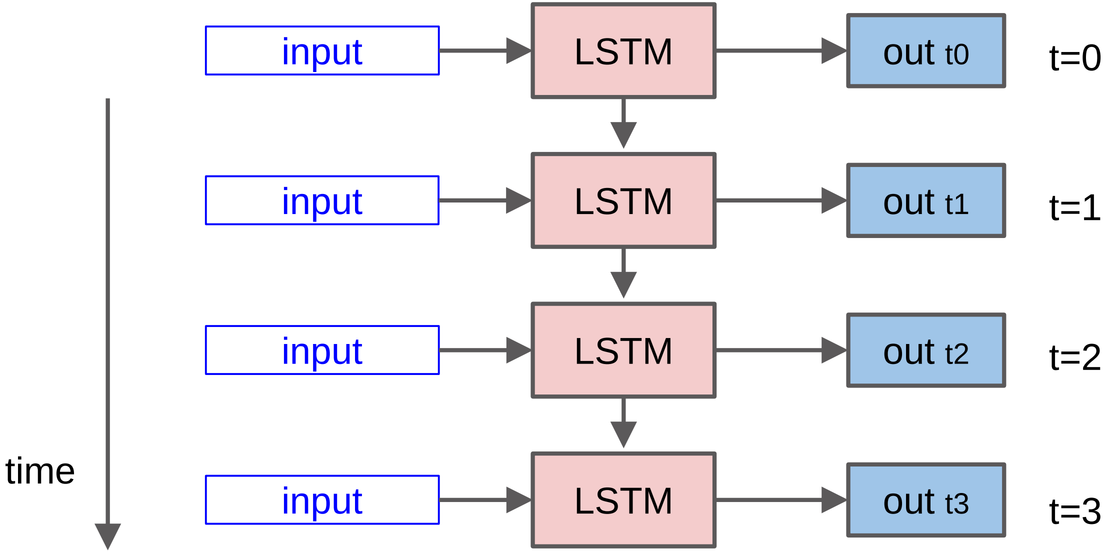
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1})$$

$$h_t = o_t \odot \phi(c_t)$$

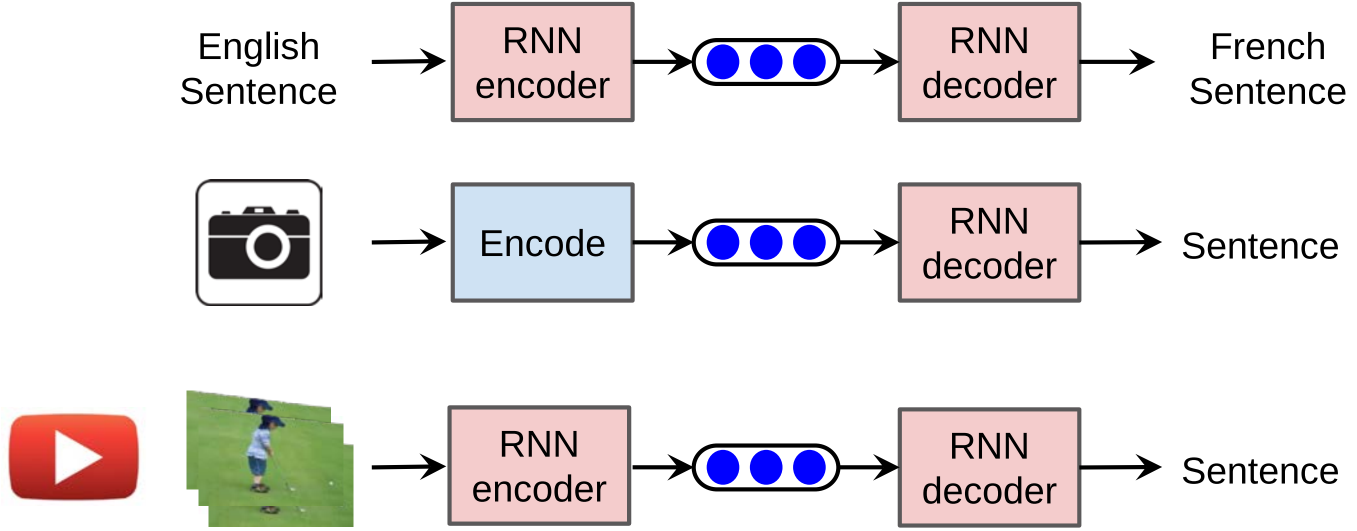
[Background] LSTM Sequence decoders

Functions are differentiable.
Full gradient is computed by backpropagating through time.
Weights updated using Stochastic Gradient Descent.

Matches state-of-the-art on:
Speech Recognition
    [Graves & Jaitly ICML'14]
Machine Translation (Eng-Fr)
    [Sutskever et al. NIPS'14]
Image-Description
    [Donahue et al. CVPR'15]
    [Vinyals et al. CVPR'15]



Key Insight:
Encode the video into hidden state vector and "decode" it to a sentence



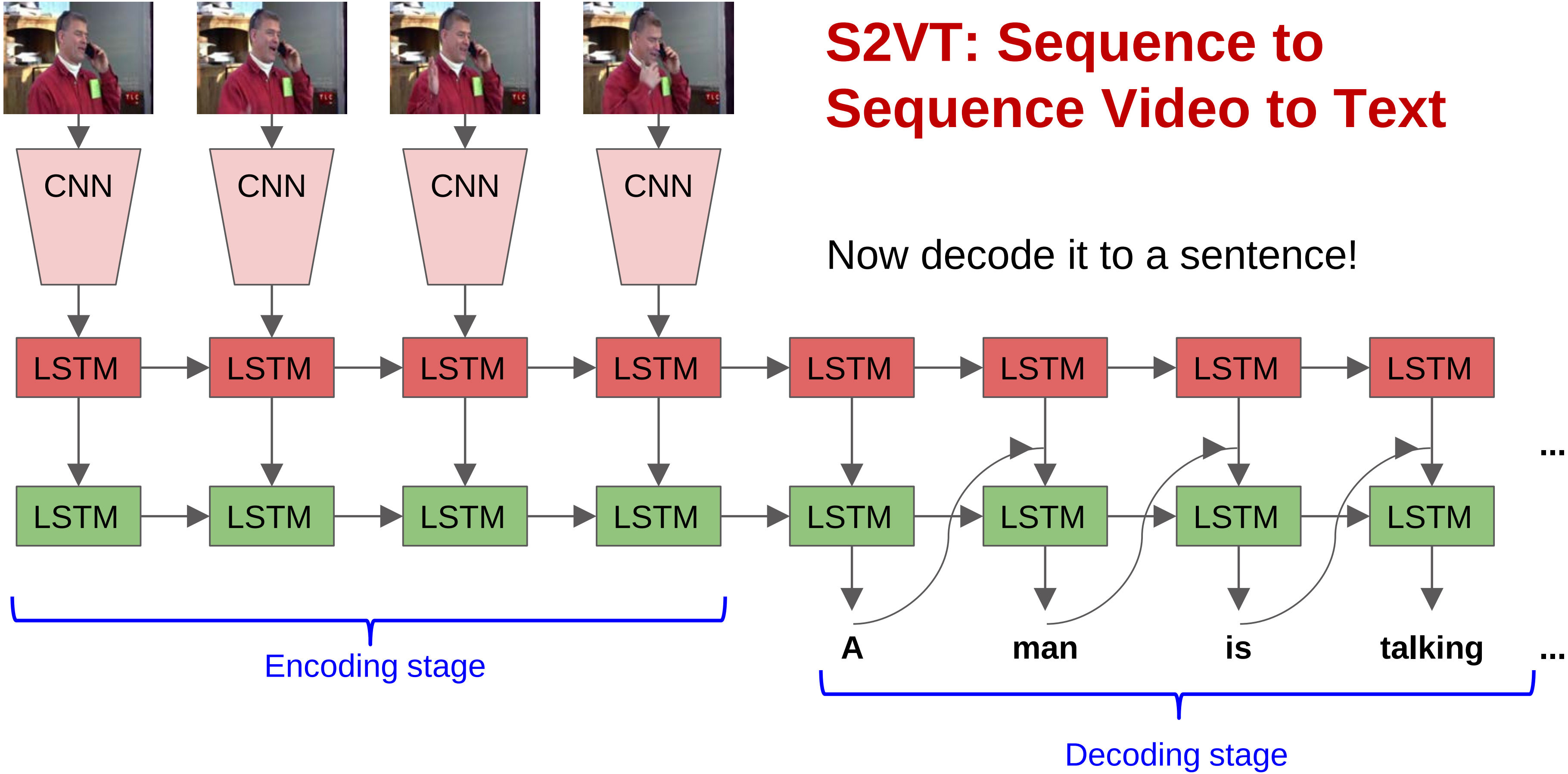
[Sutskever et al. NIPS'14]

[Donahue et al. CVPR'15]
[Vinyals et al. CVPR'15]

[Venugopalan et. al. NAACL'15]
[Venugopalan et. al. ICCV'15] (our work)

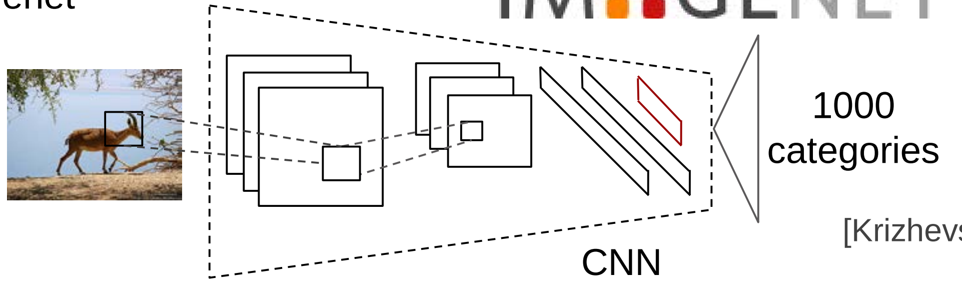
S2VT: Sequence to Sequence Video to Text

Now decode it to a sentence!



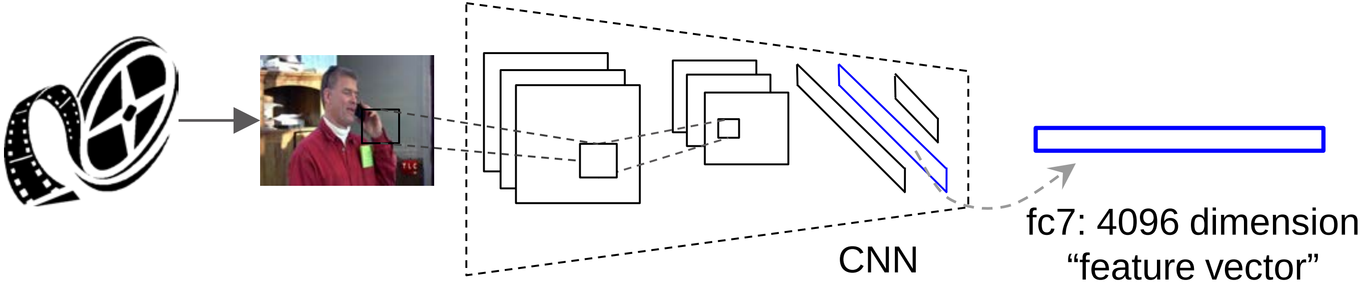
[Venugopalan et. al. ICCV’15]

1. Train on Imagenet



[Krizhevsky et al. NIPS'15]

2. Take activations from layer before classification



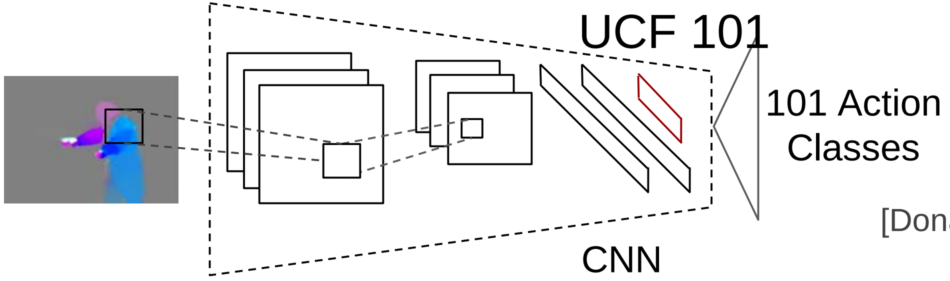
Forward propagate
Output: "fc7" features
(activations before classification layer)

Frames: RGB

Frames: Flow

Explicit Activity Recognition Features

1. Train CNN on Activity classes

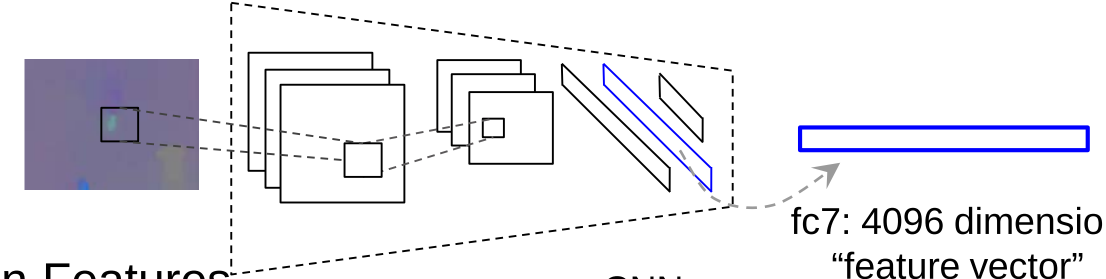


[Donahue et al. CVPR'15]

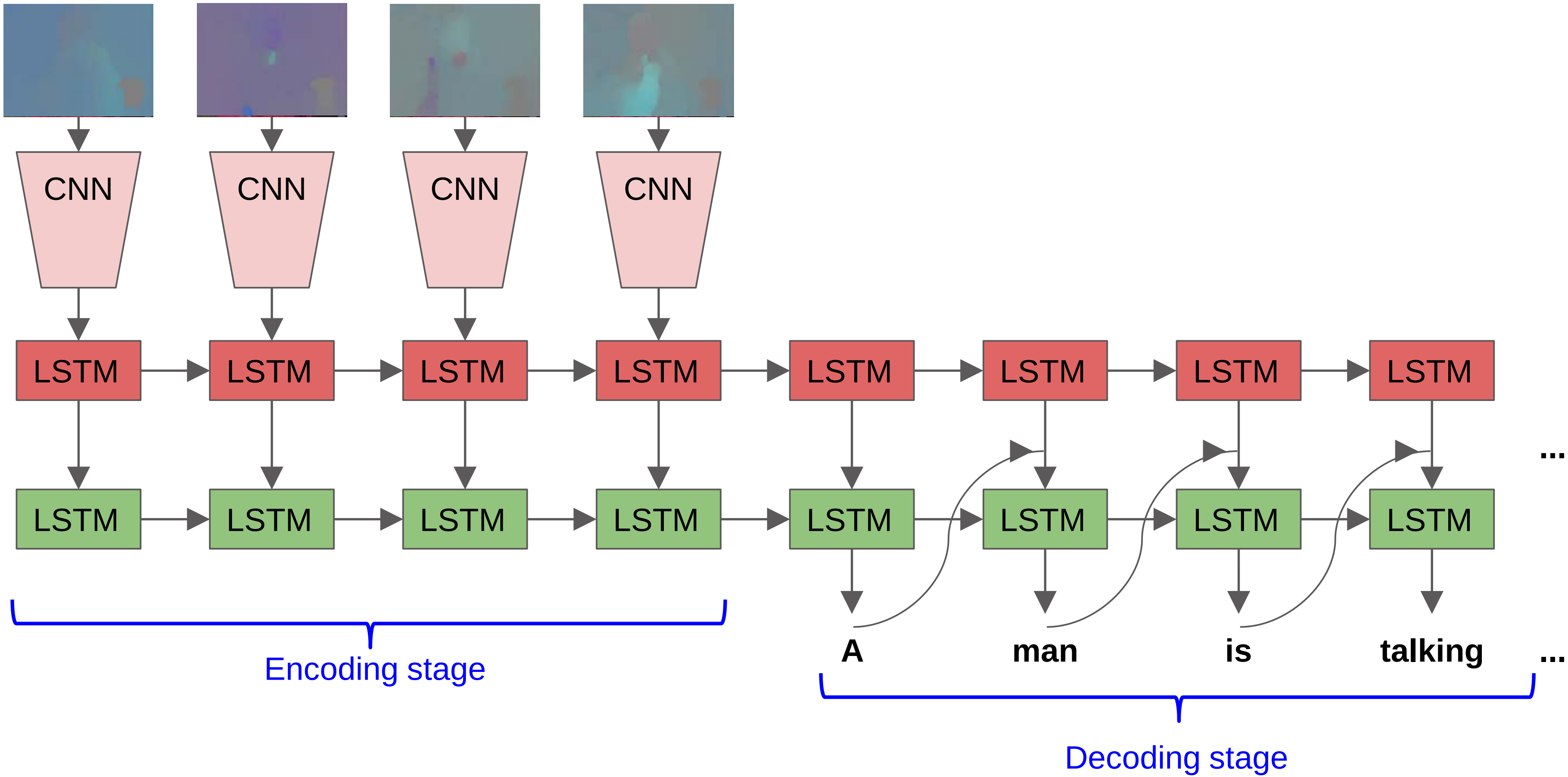
2. Use optical flow to extract flow images.

[T. Brox et. al. ECCV '04]

3. Take activations from layer before classification



Forward propagate
Output: "fc7" features
(activations before classification layer)



[Venugopalan et. al. ICCV’15]

Experiments: MSR Youtube Dataset

Microsoft Research Video Description dataset **[Chen & Dolan, ACL'11]**

Link: http://www.cs.utexas.edu/users/ml/clamp/videoDescription/

1970 YouTube video snippets

- 10-30s each
- typically single activity
- no dialogues
- 1200 training, 100 validation, 670 test

Annotations

- Descriptions in multiple languages
- ~40 English descriptions per video
- descriptions and videos collected on AMT

Youtube corpus: Sample video and gold descriptions

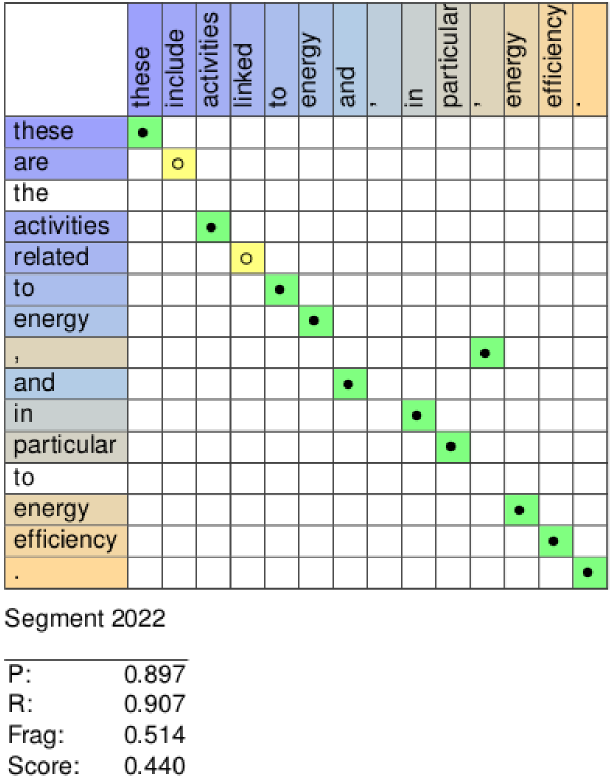
- A man appears to be **plowing** a rice field with a plow being pulled by two **oxen**.
- A team of **water buffalo** **pull** a plow through a rice paddy.
- Domesticated **livestock** are helping a man **plow**.
- A man **leads** a team of oxen down a muddy path.
- Two **oxen** **walk** through some mud.
- A man is **tilling** his land with an **ox** **pulled** plow.
- **Bulls** are **pulling** an object.
- Two **oxen** are **plowing** a field.
- The farmer is **tilling** the soil.
- A man in **ploughing** the field.

- A man is **walking** on a **rope**.
- A man is **walking** across a **rope**.
- A man is **balancing** on a **rope**.
- A man is **balancing** on a **rope** at the beach.
- A man **walks** on a **tightrope** at the beach.
- A man is **balancing** on a **volleyball net**.
- A man is **walking** on a **rope** held by poles
- A man **balanced** on a **wire**.
- The man is **balancing** on the **wire**.
- A man is **walking** on a **rope**.
- A man is **standing** in the sea shore.

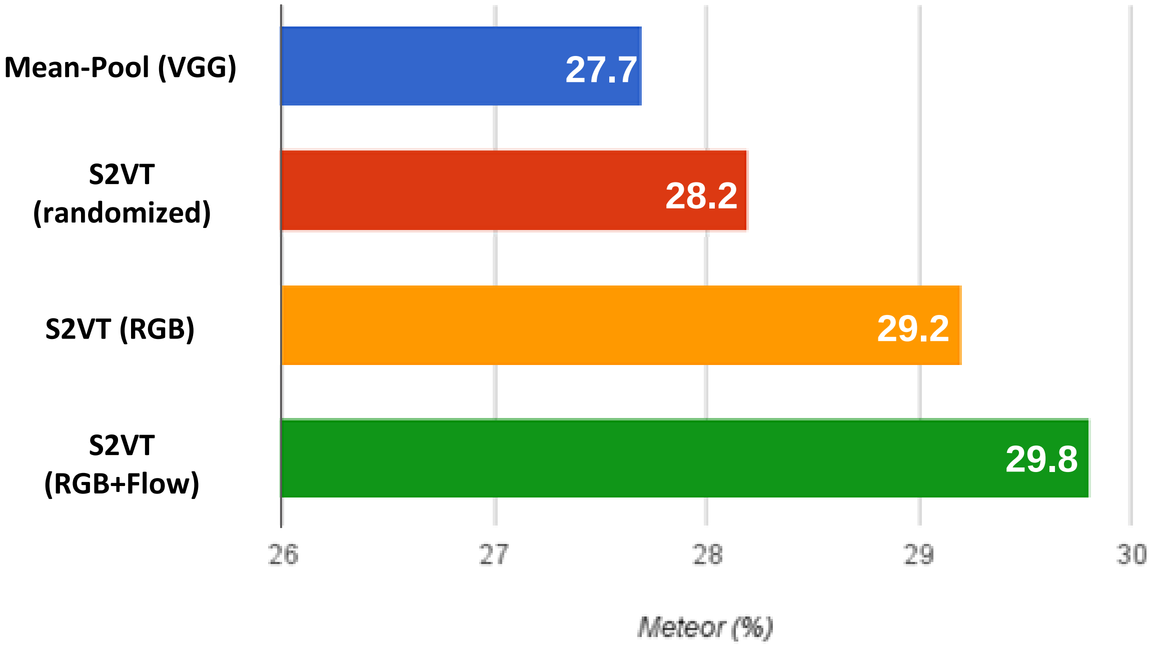
Evaluation Metric

METEOR

- scores hypotheses by aligning them to one or more reference sentences
- alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases



Results (Youtube)



Movie Corpus - DVS

CC: Queen: “Which estate?”

DVS: Looking troubled, the Queen descends the stairs. | The Queen rushes into the courtyard. She then puts a head scarf on . . . | . . . and gets into the driver’s side of a nearby Land Rover. | The Land Rover pulls away. | Three bodyguards quickly jump into a nearby car and follow her.

Processed:
Looking troubled, someone descends the stairs.

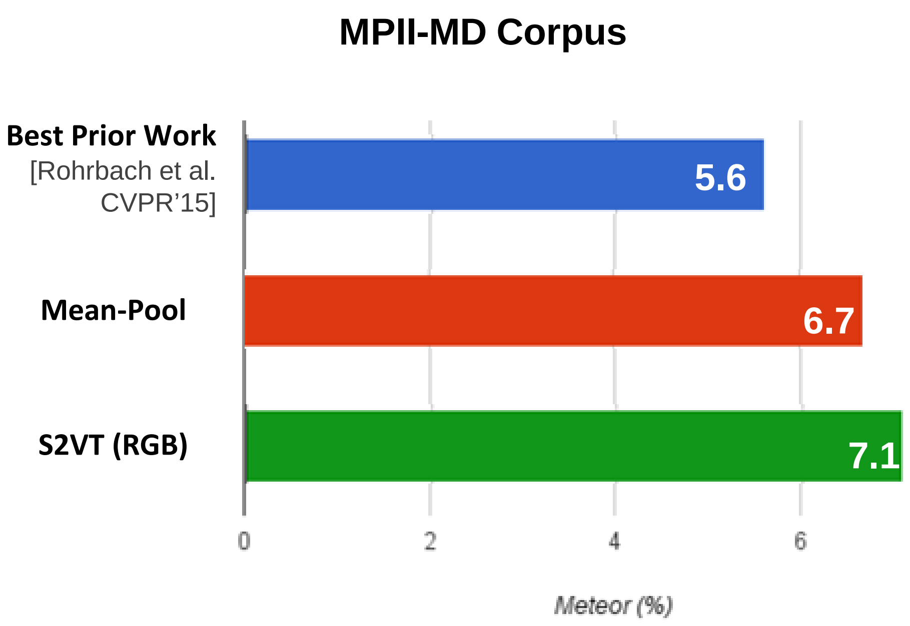
Someone rushes into the courtyard. She then puts a head scarf on ...

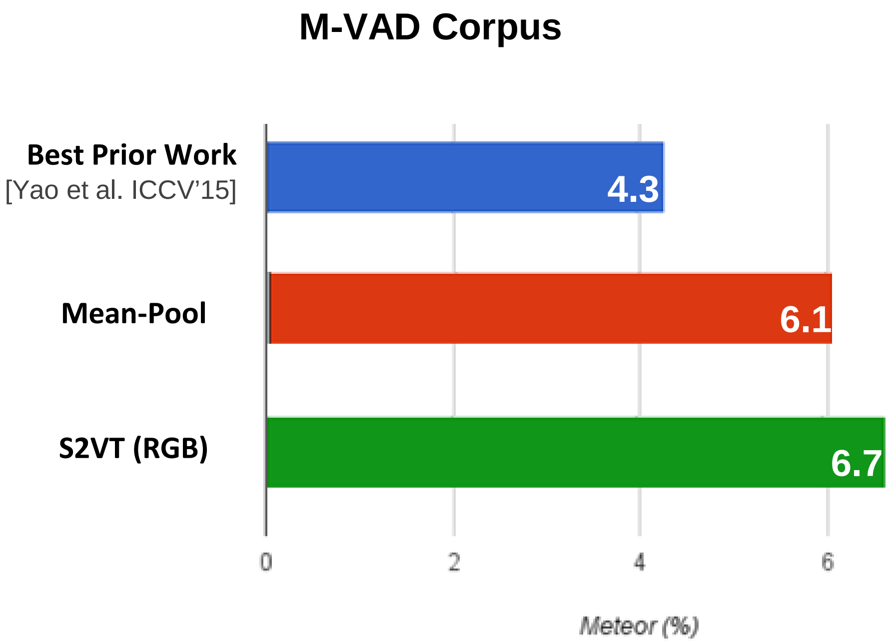
Evaluation: Movie Corpora

MPII-MD
- MPII, Germany
- DVS alignment: semi-automated and crowdsourced
- 94 movies
- 68,000 clips
- Avg. length: 3.9s per clip
- **~1 sentence per clip**
- 68,375 sentences

M-VAD
- Univ. of Montreal
- DVS alignment: semi-automated and crowdsourced
- 92 movies
- 46,009 clips
- Avg. length: 6.2s per clip
- **1-2 sentences per clip**
- 56,634 sentences

Results (M-VAD Movie Corpus)





Examples (M-VAD Movie Corpus)

MPII-MD: https://youtu.be/XTq0huTXj1M
M-VAD: https://youtu.be/pER0mjzSYaM

Today

ICCV15 – end-to-end video captioning

ACM MM16 – multimodal video captioning

CVPR17 – caption-guided video saliency

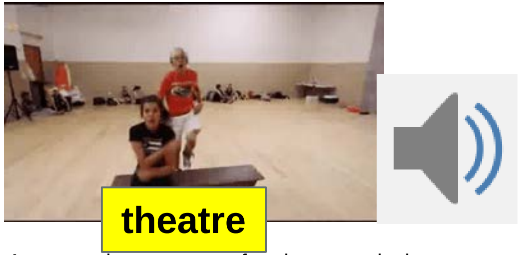
Multimodal Video Description

Vasili Ramanishka[1], Abir Das[1], Dong Huk Park[3], Subhashini Venugopalan[2], Lisa Anne Hendricks[3], Marcus Rohrbach[3], Kate Saenko[1]

[1] Boston University, MA
[2] University of Texas Austin, TX
[3] UC Berkeley, CA

Problem: how to incorporate non-visual information?

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.



1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical.
3. Dancers are playing a routine.
4. People are dancing in a mus
5. Some people are acting anc
performance.

singing

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is … of
mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

making a speech

1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of speculators.

Xu et al., CVPR 2016

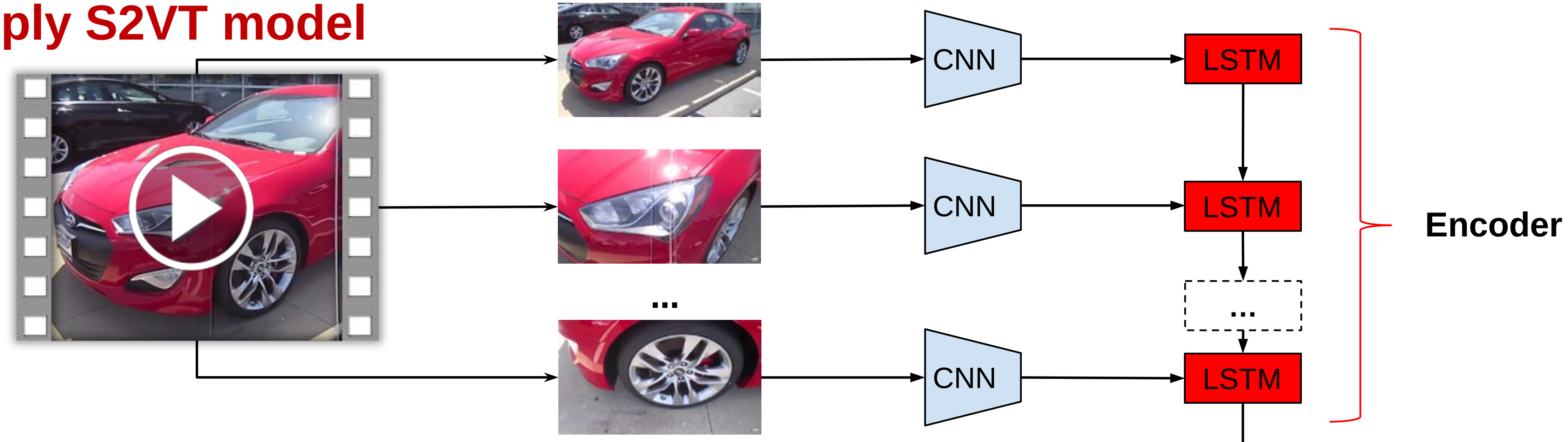
MSR-VTT Dataset

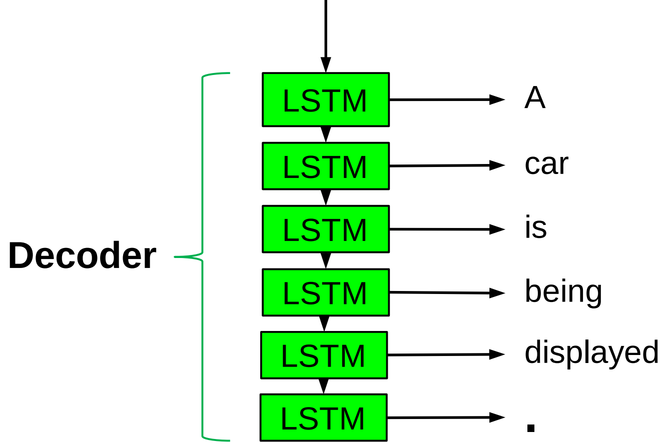
1. A white car is drifting.
2. Cars racing on a road surrounded by lots of people.
3. Cars are racing down a narrow road.
4. A race car races along a track.
5. A car is drifting in a fast speed.

Dataset	Context	Sentence Source	#Video	#Clip	#Sentence	#Word	Vocabulary	Duration (hrs)
YouCook [5]	cooking	labeled	88	–	2,668	42,457	2,711	2.3
TACos [25, 28]	cooking	AMT workers	123	7,206	18,227	–	–	–
TACos M-L [26]	cooking	AMT workers	185	14,105	52,593	–	–	–
M-VAD [32]	movie	DVS	92	48,986	55,905	519,933	18,269	84.6
MPII-MD [27]	movie	DVS+Script	94	68,337	68,375	653,467	24,549	73.6
MSVD [3]	multi-category	AMT workers	–	1,970	70,028	607,339	13,010	5.3
MSR-VTT-10K	20 categories	AMT workers	7,180	10,000	200,000	1,856,523	29,316	41.2

Xu et al., CVPR 2016

Apply S2VT model



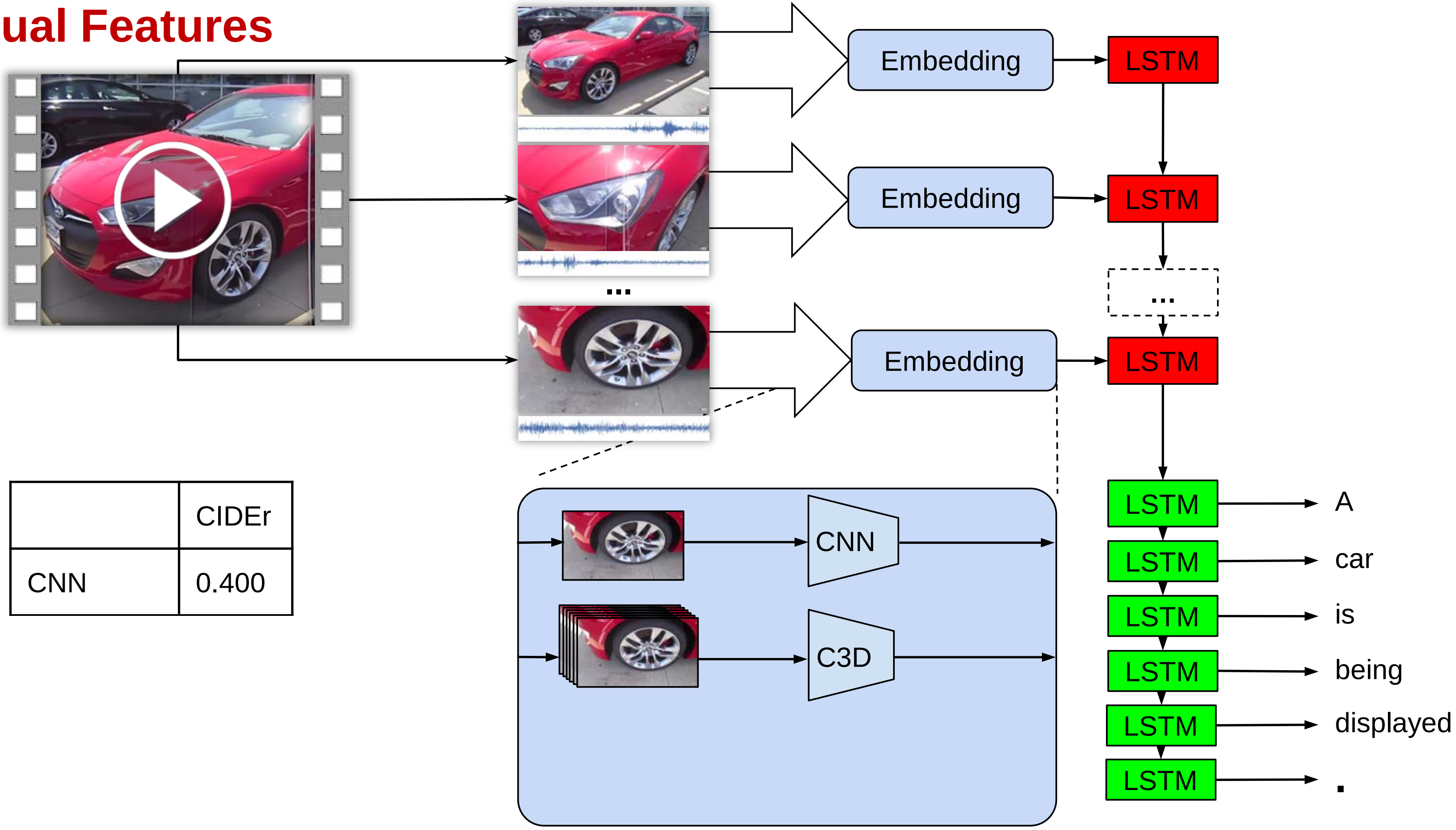


Start with S2VT model

- Simplify to single LSTM layer
- Train end-to-end on MSR-VTT

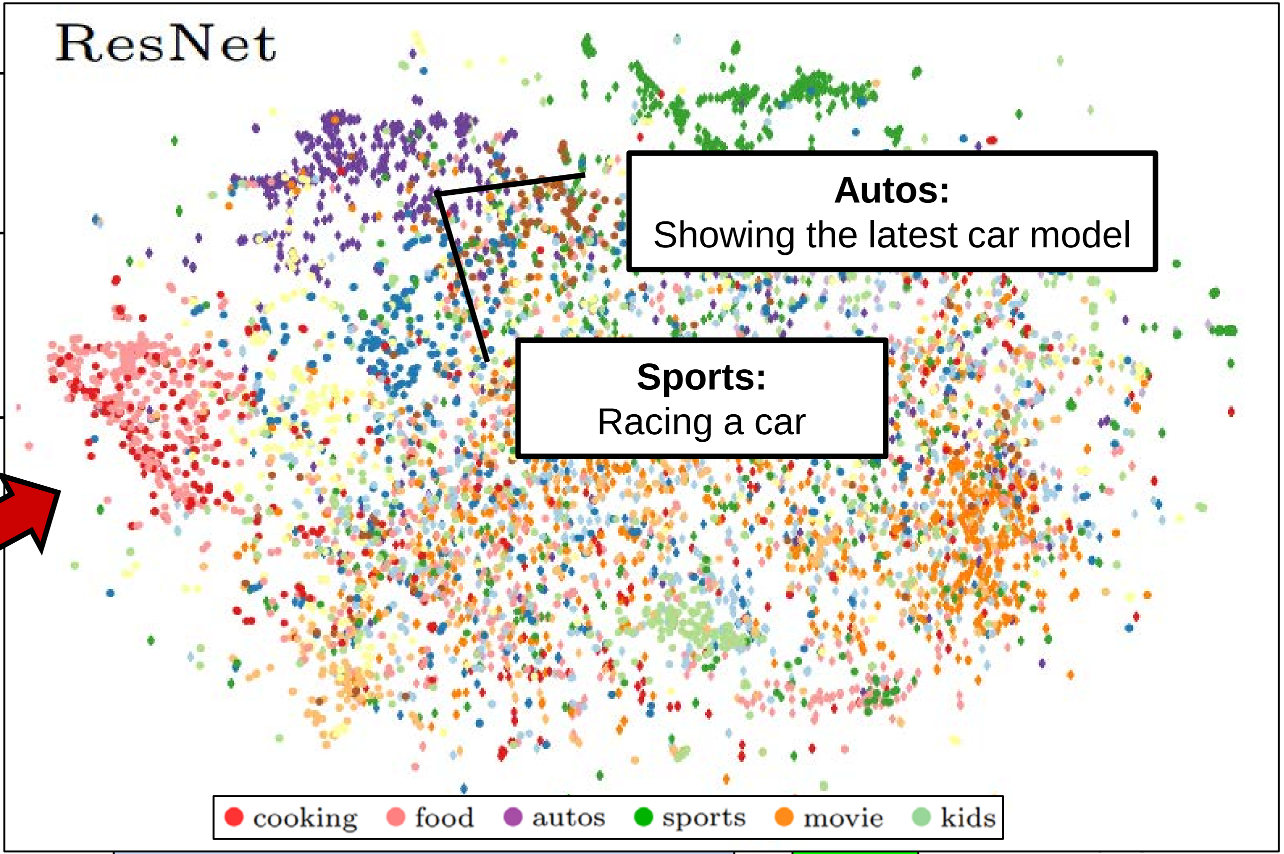
Venugopalan et al., ICCV 2015

Visual Features



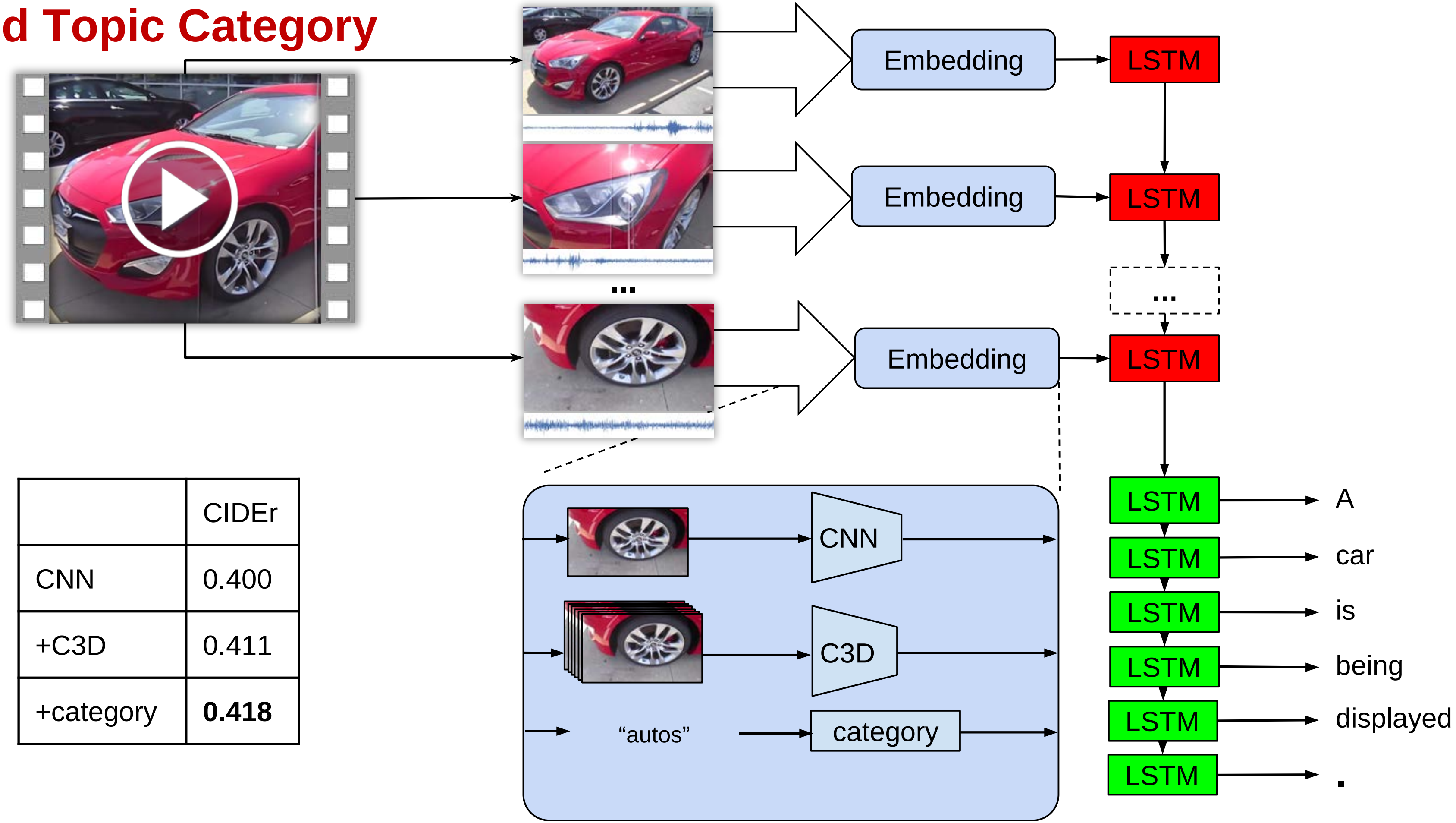
	CIDEr
CNN	0.400

Add Topic Category



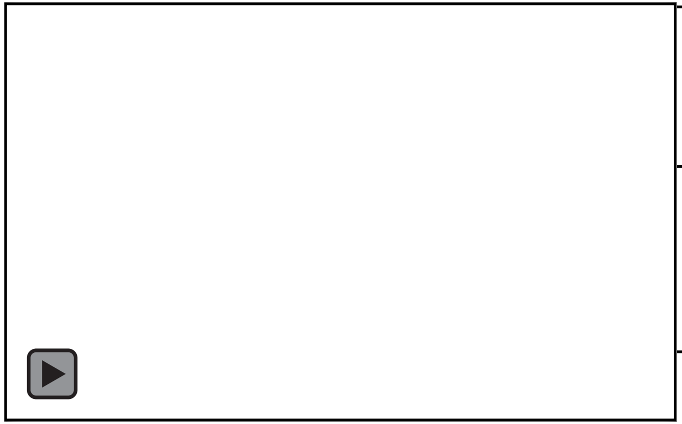
	CIDEr
CNN	0.400
+C3D	0.411

Add Topic Category

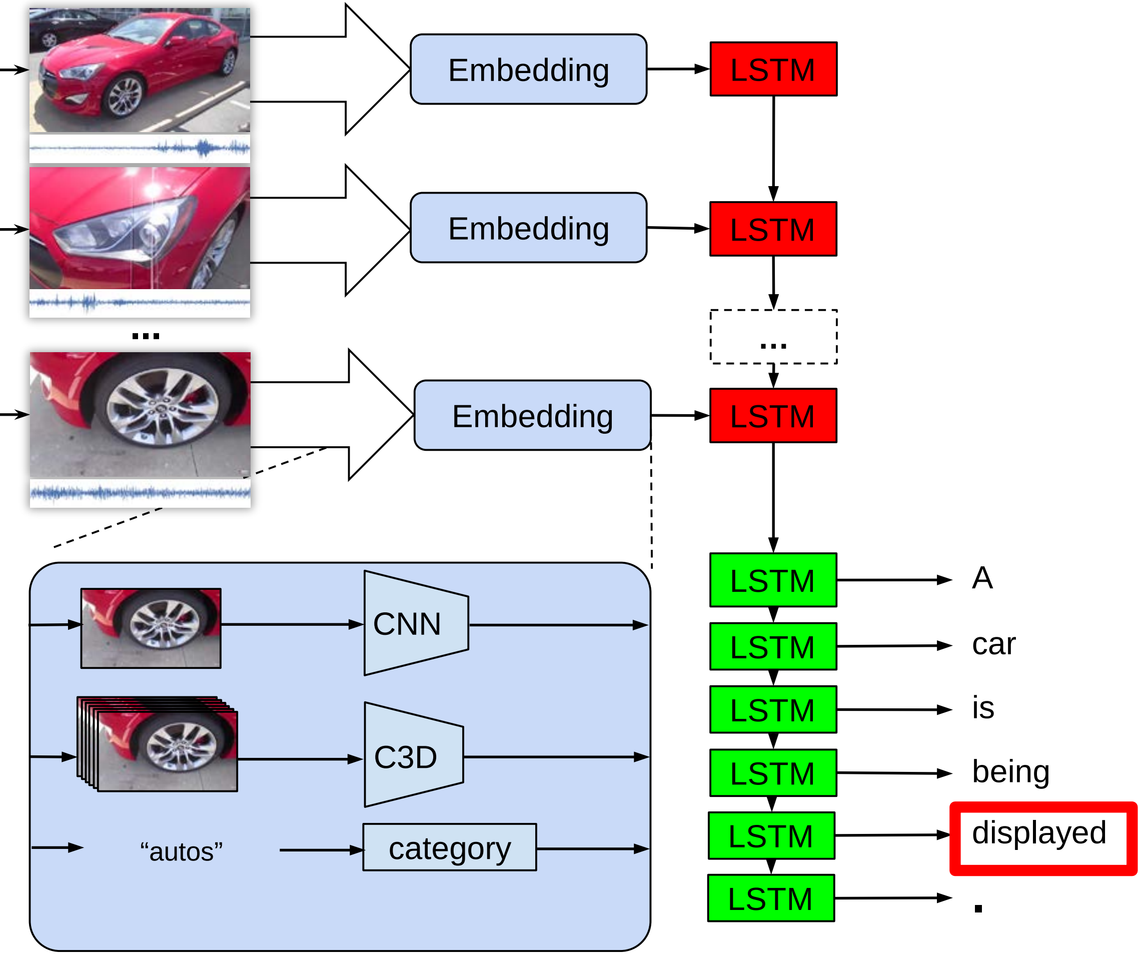


	CIDEr
CNN	0.400
+C3D	0.411
+category	**0.418**

Add Sound Features

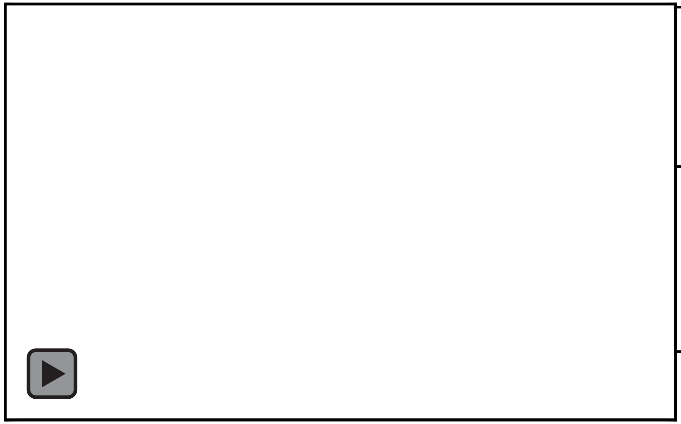


No audio



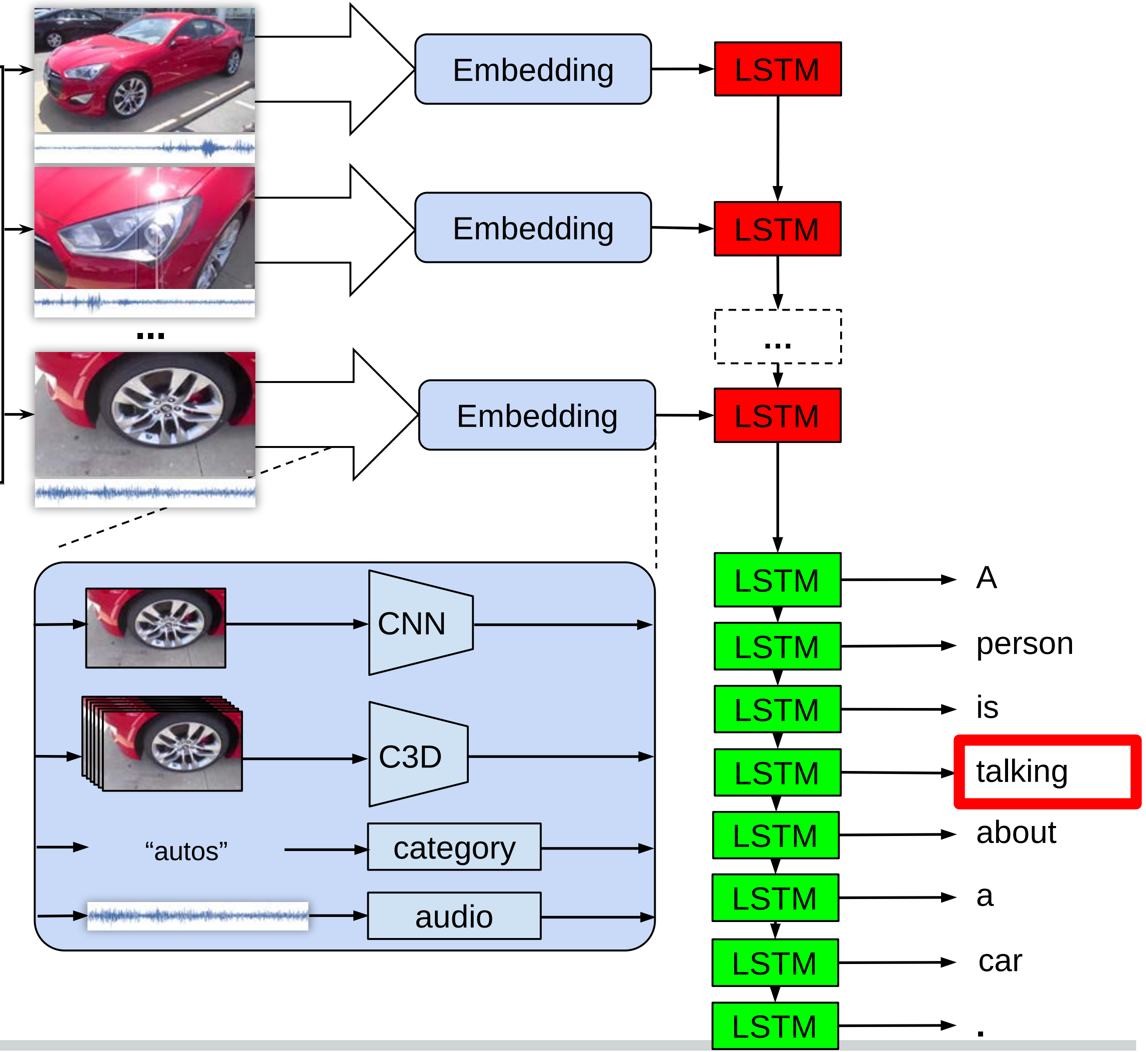
	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418

Add Sound Features

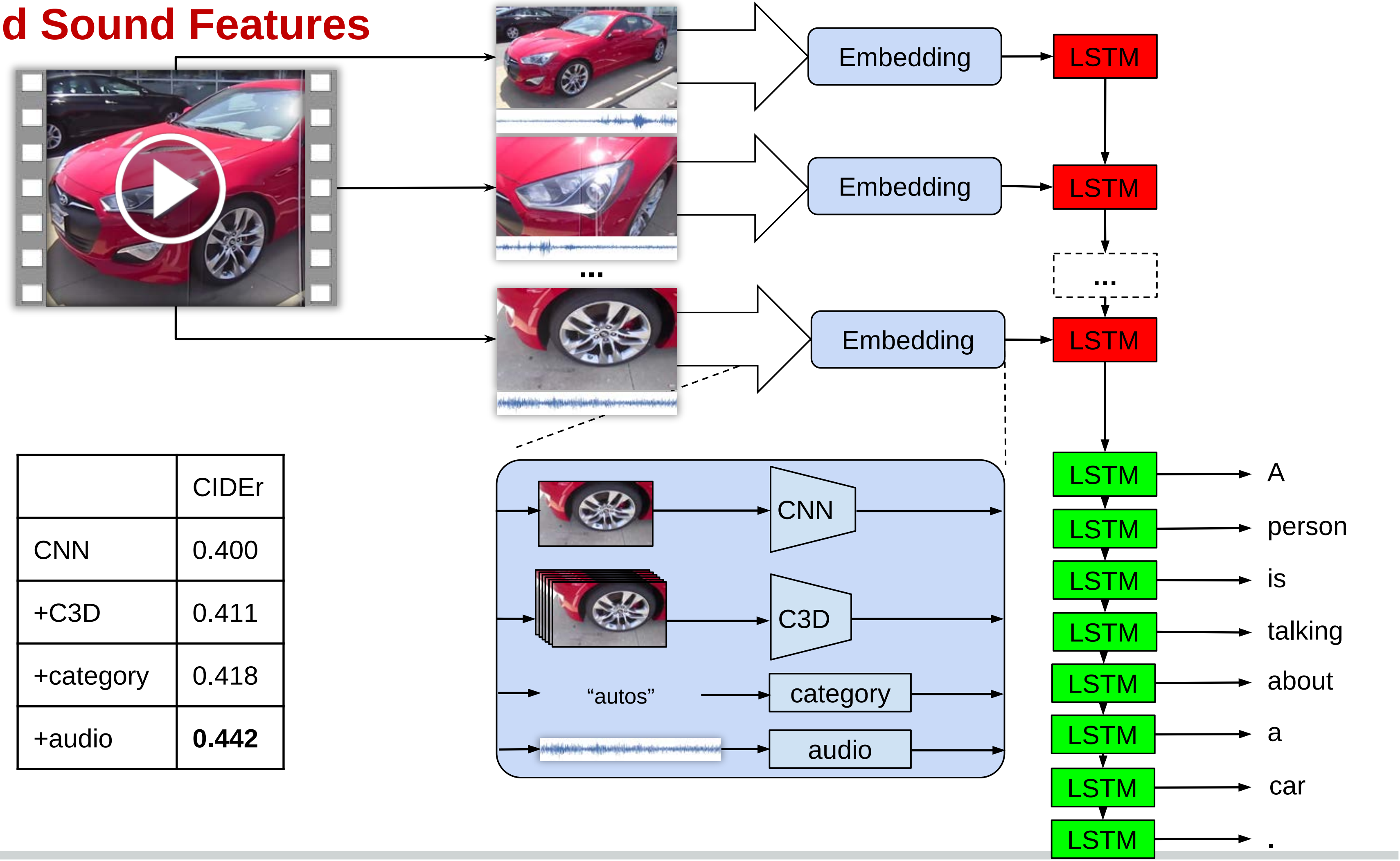


With audio

	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418

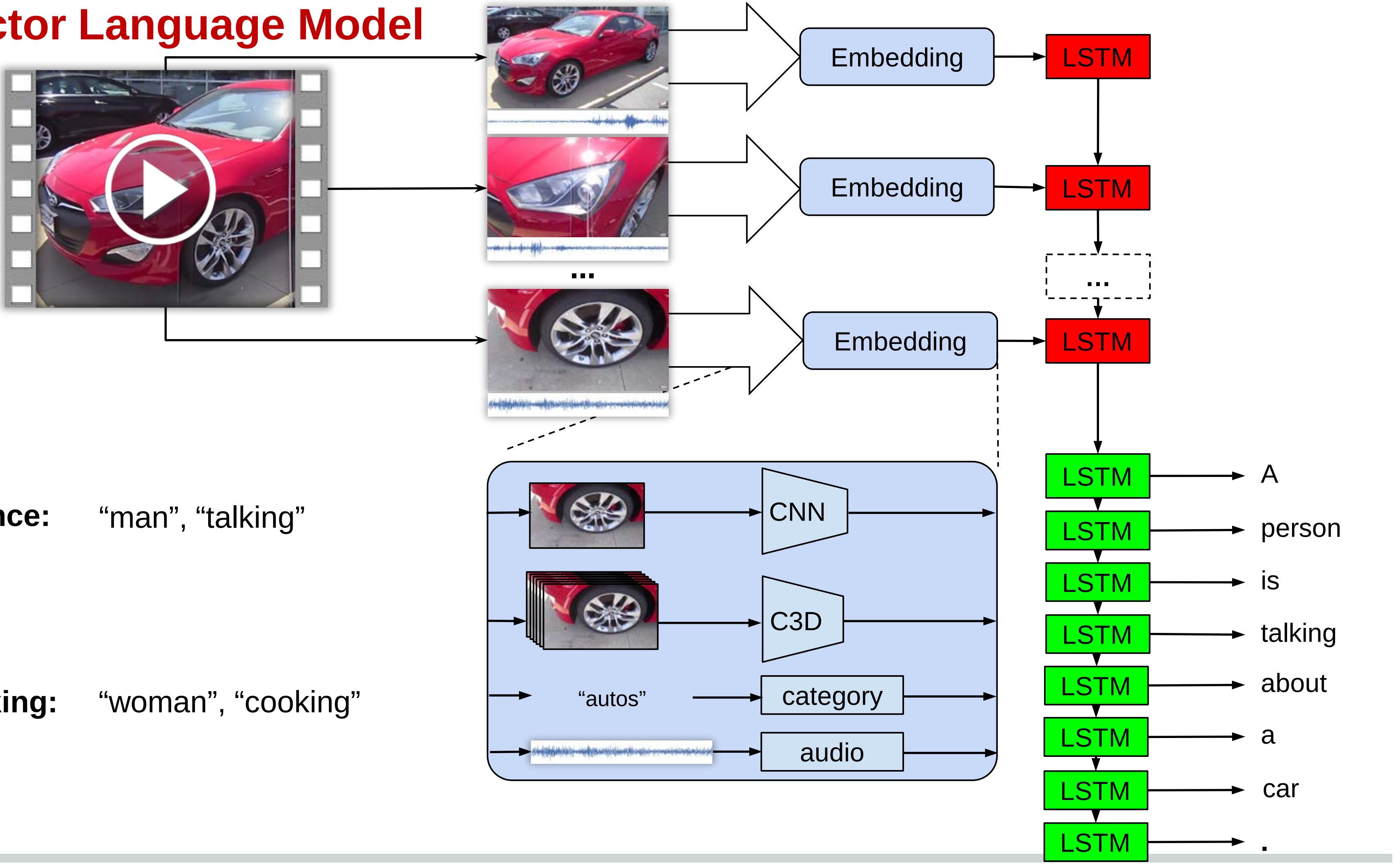


Add Sound Features



	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418
+audio	**0.442**

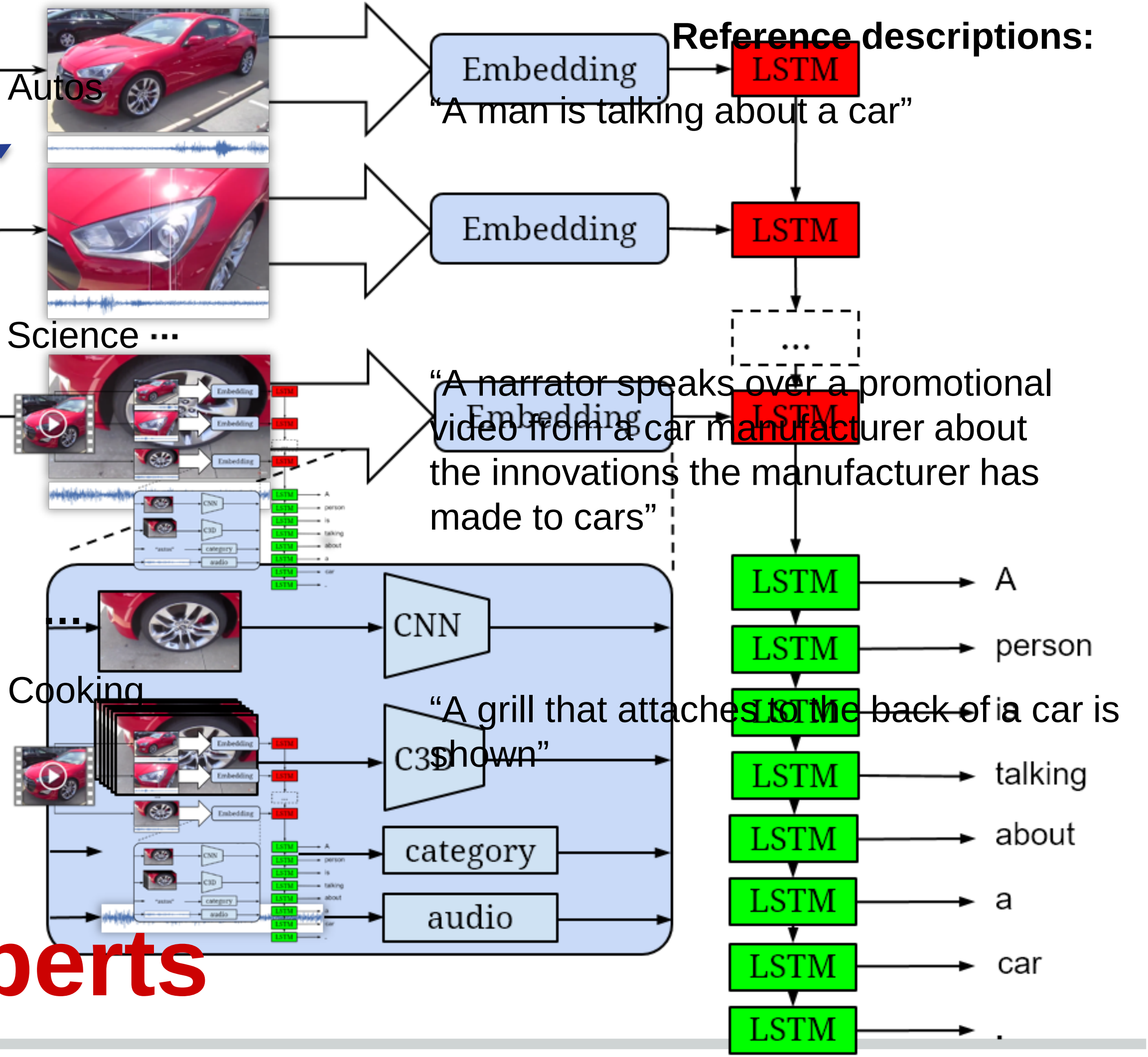
Factor Language Model



Science: "man", "talking"

Cooking: "woman", "cooking"

Factor Language Model



Reference descriptions:

"A man is talking about a car"

"A narrator speaks over a promotional video from a car manufacturer about the innovations the manufacturer has made to cars"

"A grill that attaches to the back of a car is shown"

	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418
+audio	0.442
experts	**0.465**

Network of experts

Our final model

	CIDEr
CNN	0.400
+C3D	0.411
+category	0.418
+audio	0.442
experts	**0.465**

- Baseline model

  → **Encoder – decoder approach (S2VT)**
- Capture activities and motion

  → **C3D as motion features**
- Capture sound and audio

  → **MFCC as audio features**
- Topic aware model to capture language differences

  → **Network of experts**

ACM MM 2016 Video Description Challenge

Automatic evaluation

Rank	Team	Organization	BLEU@4	Meteor	CIDEr-D	ROUGE-L
1	v2t_navigator	RUC & CMU	0.408	0.282	0.448	0.609
2	Aalto	Aalto University	0.398	0.269	0.457	0.598
3	**VideoLAB**	UML & Berkeley & UT-Austin	0.391	0.277	0.441	0.606
…						
21						

ACM MM 2016 Video Description Challenge

Human evaluation

Best on “relevance” as judged by humans

Rank	Team	Organization	**Coherence**	**Relevance**	**Helpful for blind**
1	Aalto	Aalto University	3.263	3.104	3.244
2	v2t_navigator	RUC & CMU	3.261	3.091	3.154
3	**VideoLAB**	UML & Berkeley & UT-Austin	3.237	**3.109**	3.143
			...		
21					

Today

ICCV15 – end-to-end video captioning

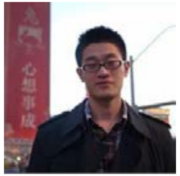
ACM MM16 – multimodal video captioning

CVPR17 – caption-guided video saliency

Top-down saliency guided by captions


Vasili
Ramanishka
Boston University


Abir
Das
Boston University


Jianming
Zhang
Adobe Research

Explaining the network's captions

Predicted sentence: A woman is cutting a piece of meat

can the network localize objects?

Neural Attention Models

"Attention": Sequentially processes regions in a single image.
Objective: Model learns "where to look" next.

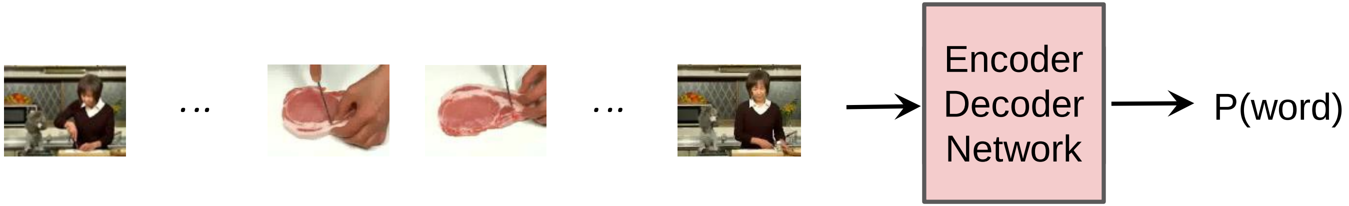
Image Captioning

girl teddy bear

Show, Attend and Tell
[Xu et al. ICML'15]

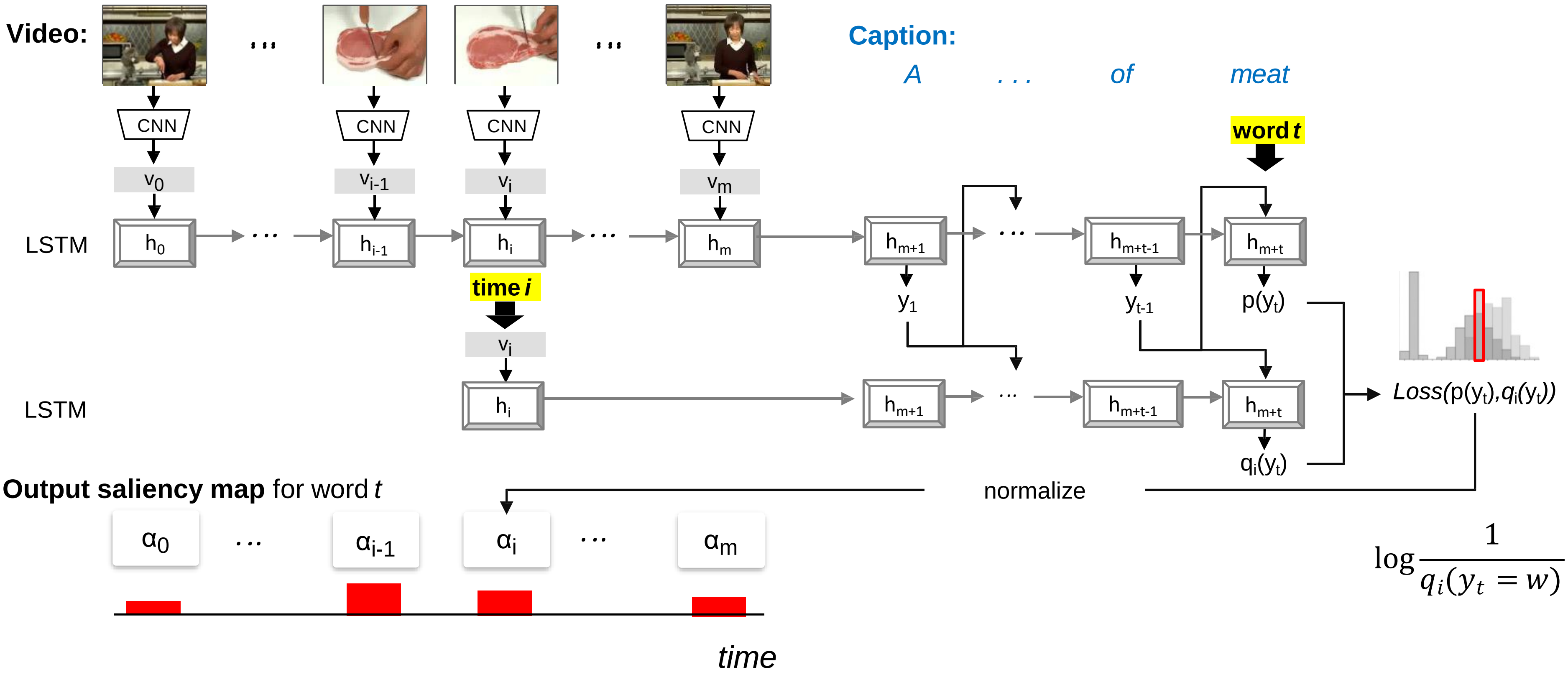
- *soft attention* adds special attention layer
- Only spatial or only temporal
- Can we get spatio-temporal attention?

Key idea: probe the network with small part of input, look at change in prob(word)

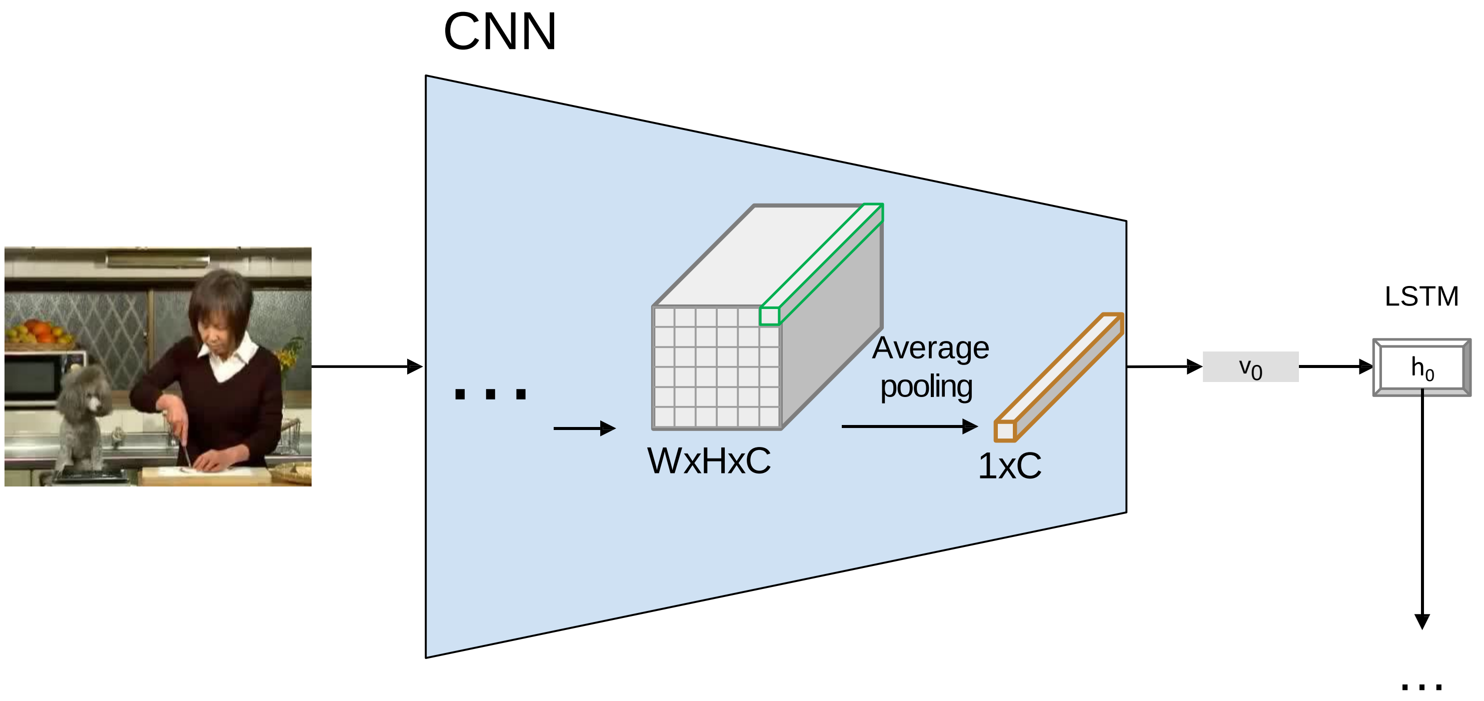


- *No need for* special attention layer
- Get spatio-temporal attention for free

Approach: temporal saliency

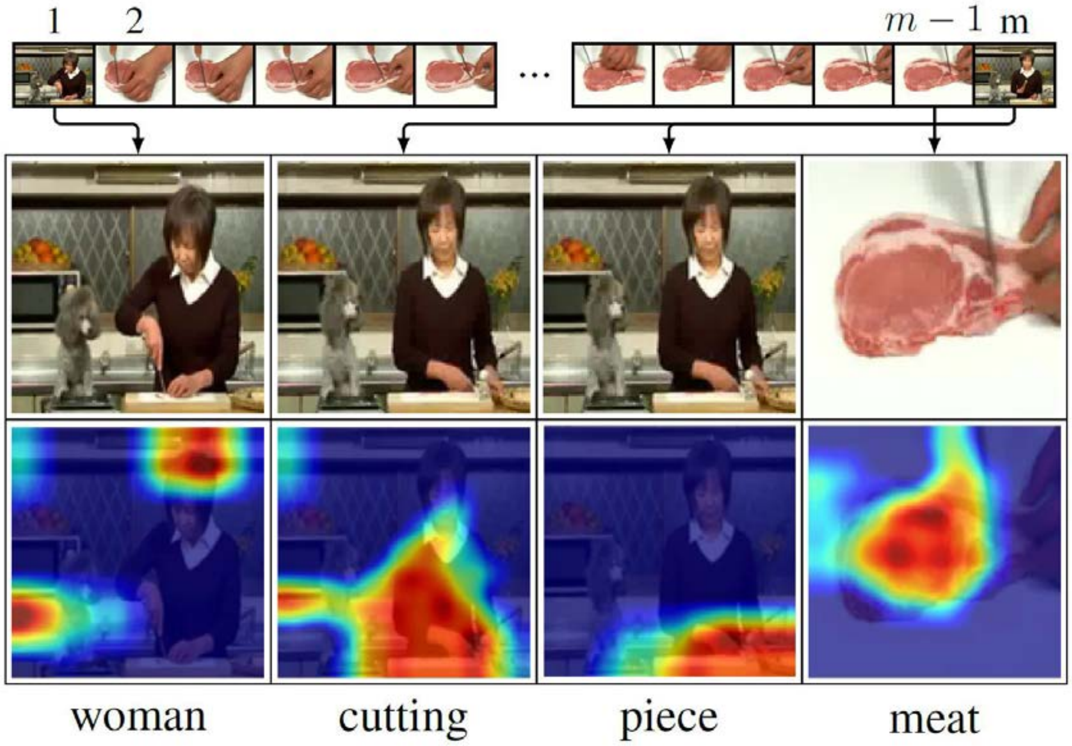


Spatial localization (almost) for free



Spatiotemporal saliency

Predicted sentence: A **woman** is **cutting** a **piece** of **meat**



Spatiotemporal saliency

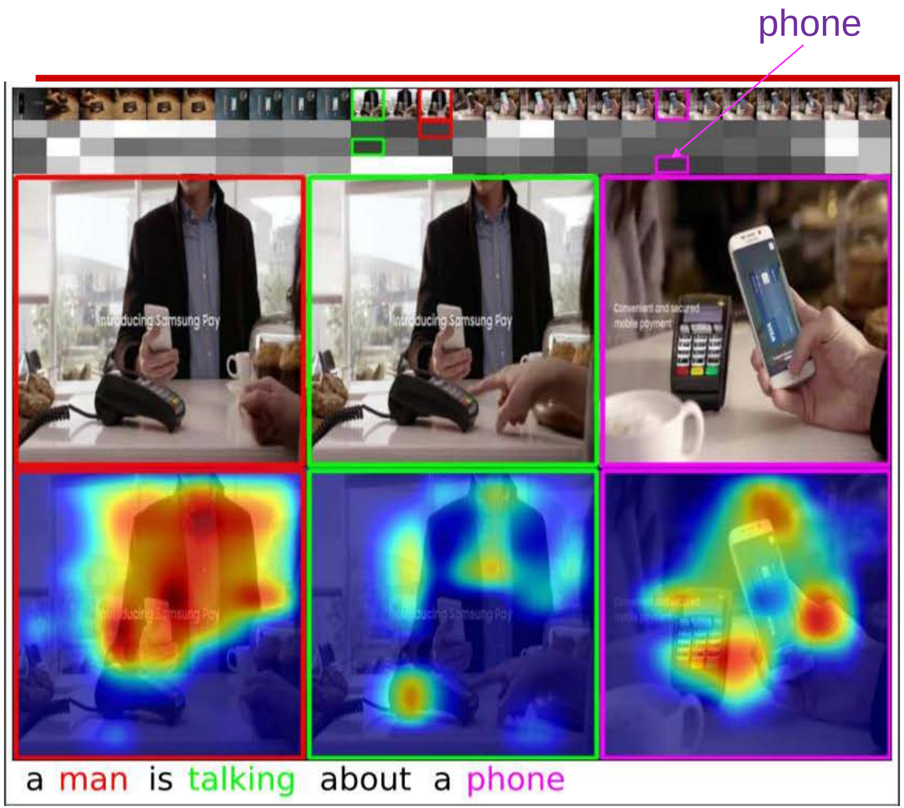


Image captioning with the same architecture

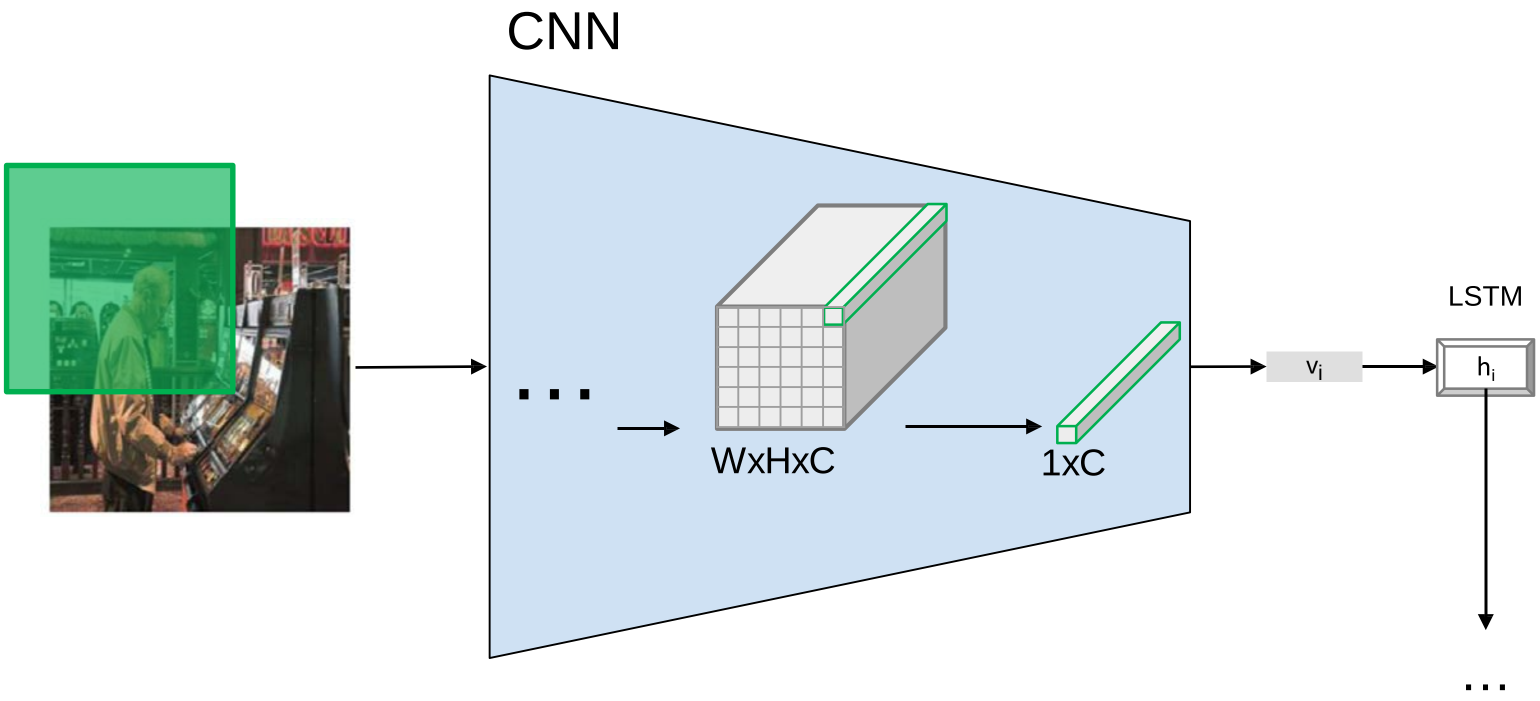
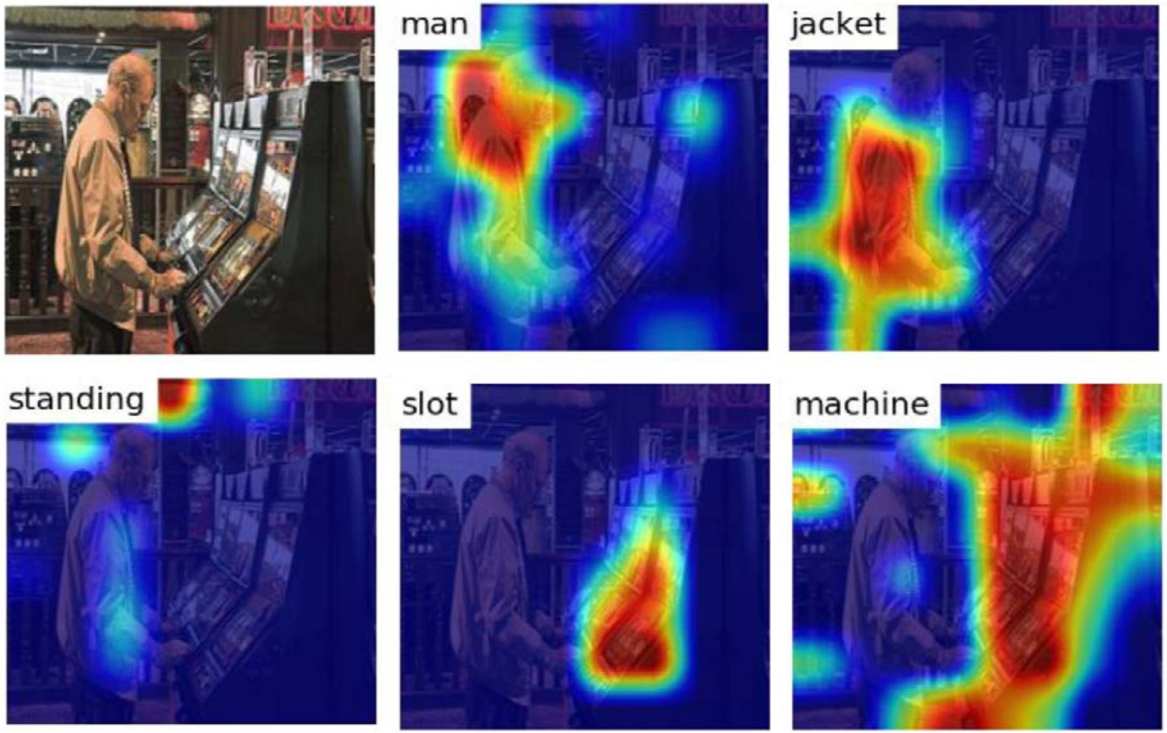
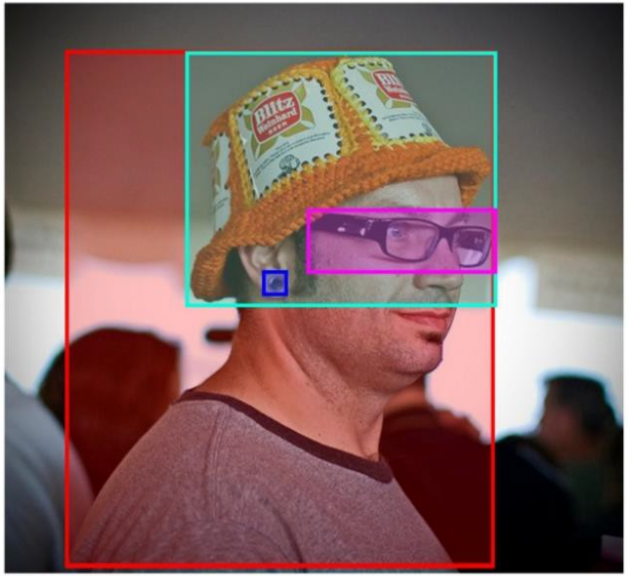


Image captioning with the same architecture

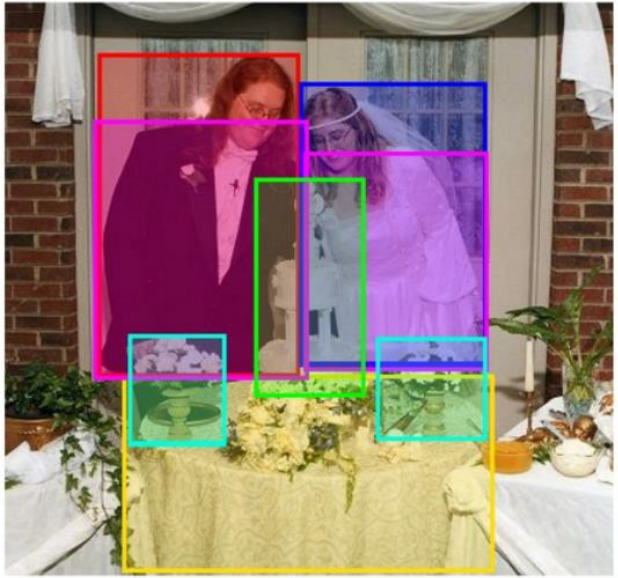
Input query: A **man** in a **jacket** is **standing** at the **slot machine**



Flickr30kEntities



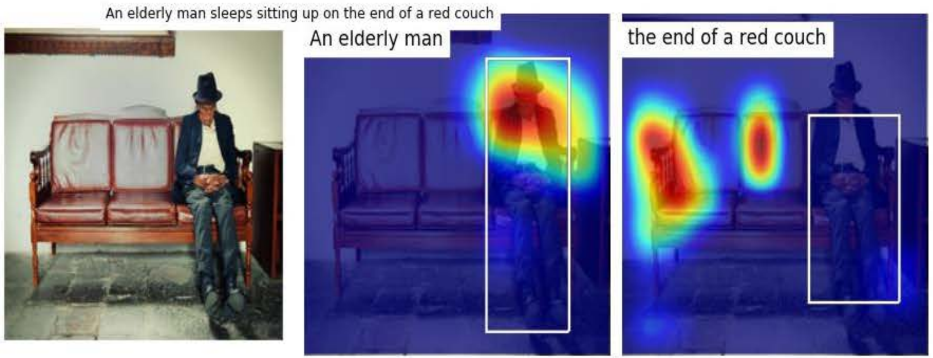
A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

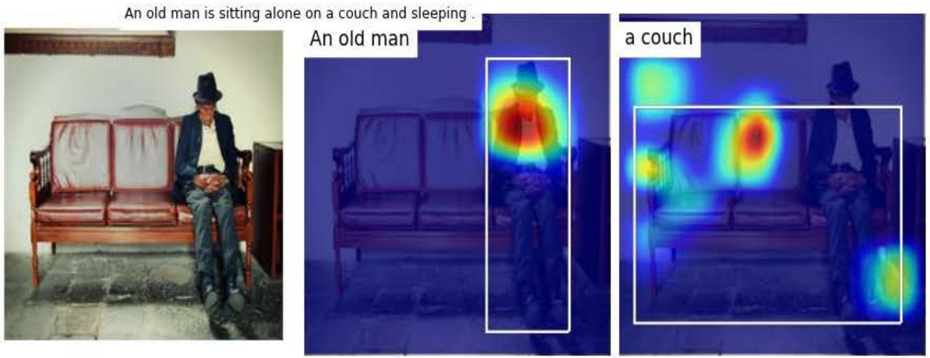


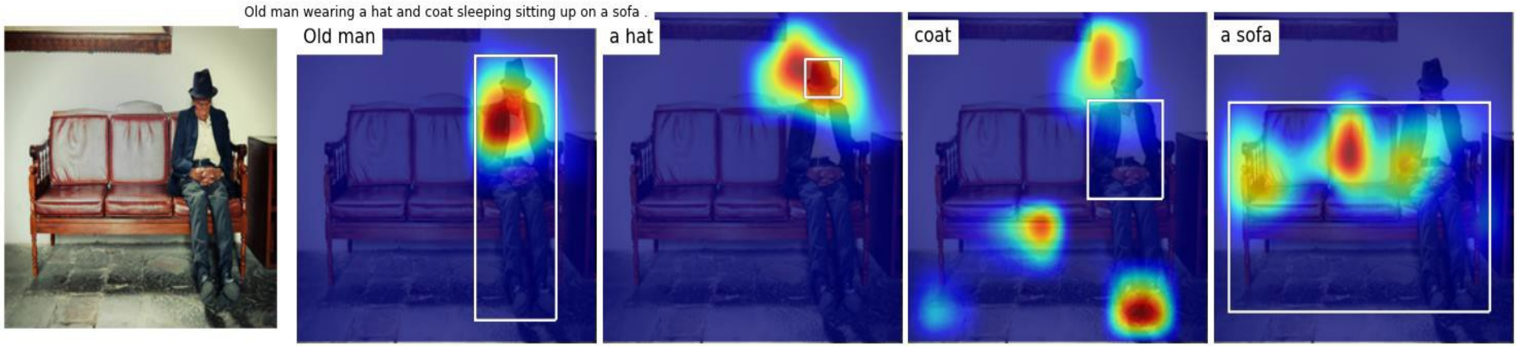
A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Plummer et al., ICCV 2015

Pointing game in Flickr30kEntities







Flickr30kEntities

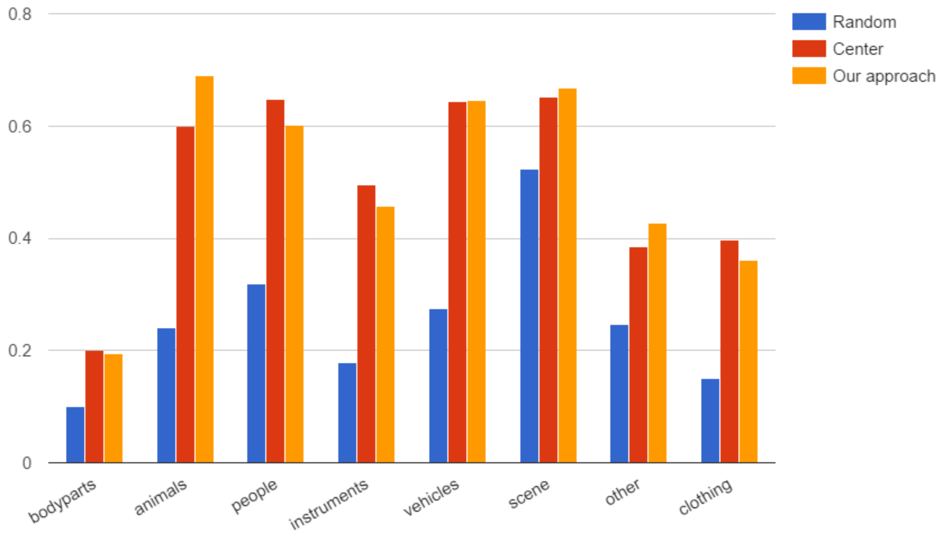
Attention correctness

	Avg per NP
Baseline [14]	0.321
SA [14]	0.387
SA-supervised [14]	0.433
Baseline*	0.325
Our model	**0.473**

Captioning performance

Model	Dataset	METEOR [9]
Soft-Attn [28]	MSVD	30.0
Our Model	MSVD	31.0
Soft-Attn [12]	MSR-VTT	25.4
Our Model	MSR-VTT	25.9
Soft-Attn [27]	Flickr30k	18.5
Our Model	Flickr30k	18.3

Pointing game accuracy

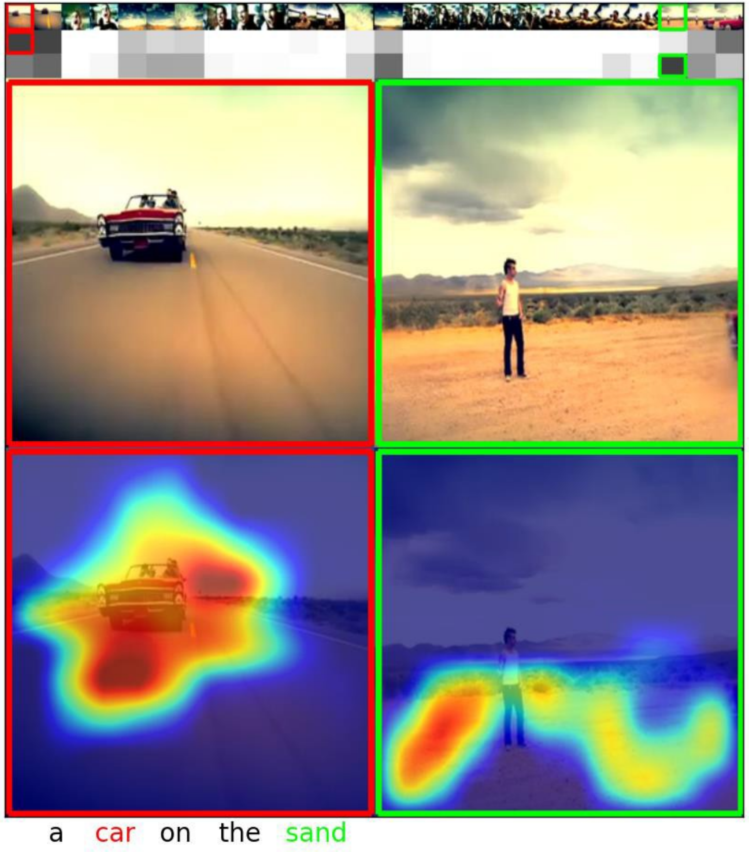


[14] C. Liu, J. Mao, F. Sha, and A. L. Yuille. Attention correctness in neural image captioning, 2016, implementation of K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML 2015

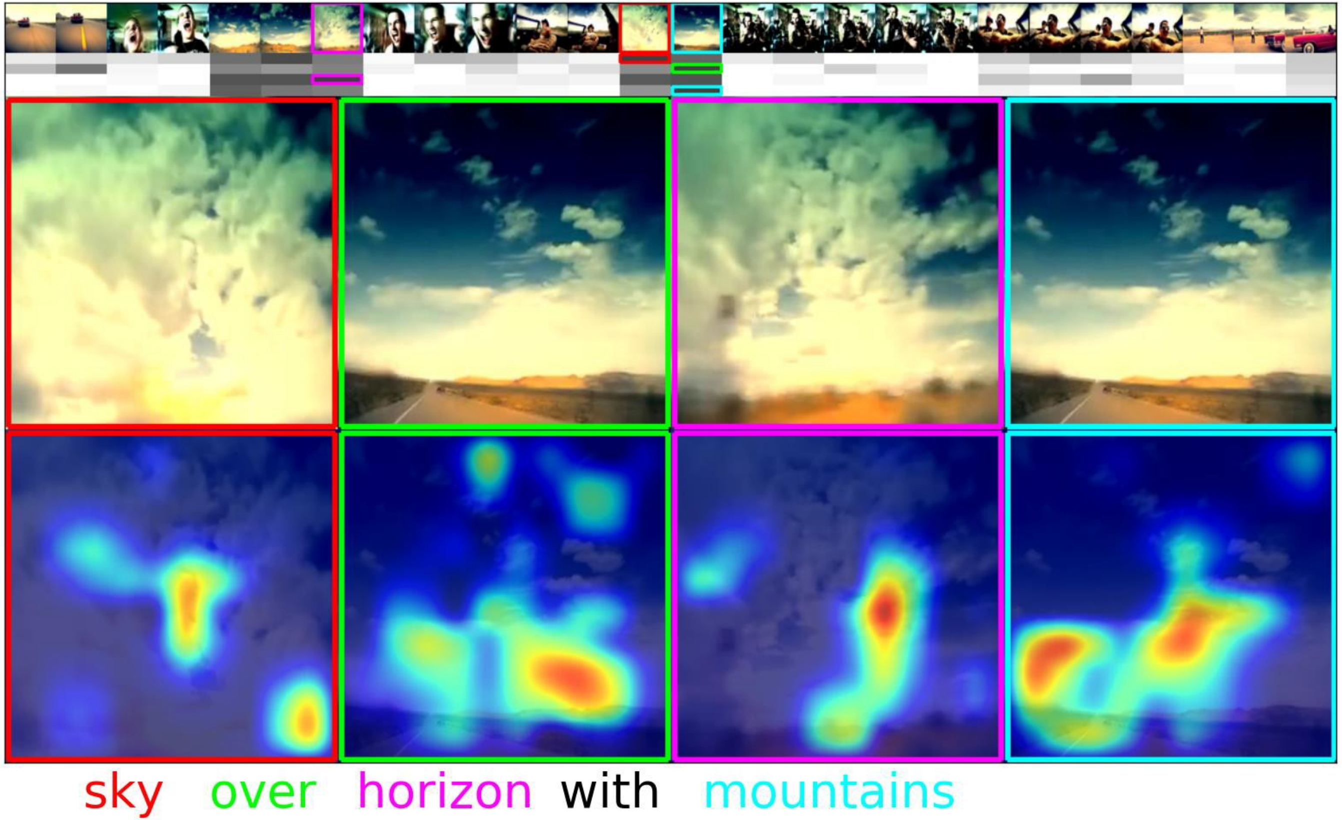
Video summarization: predicted sentence

Video summarization: arbitrary query

Video summarization: arbitrary query

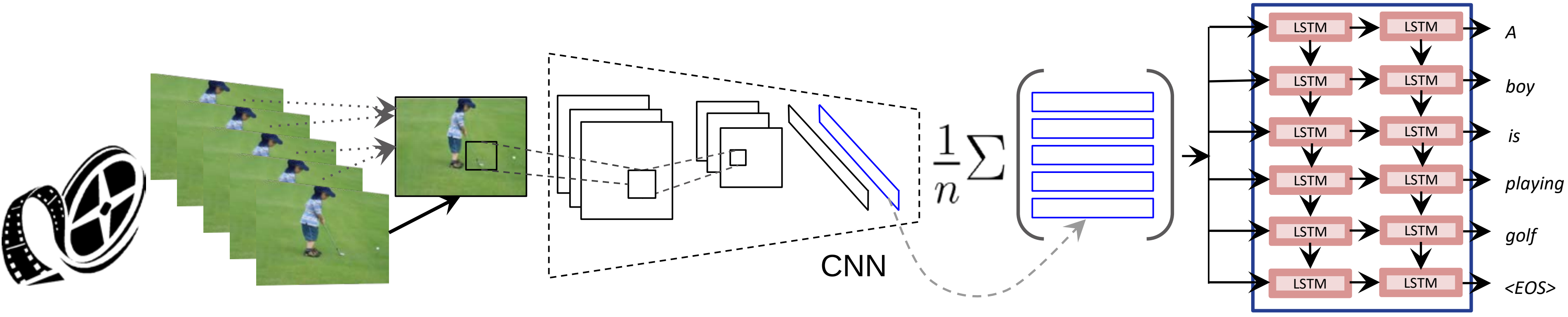


Video summarization: arbitrary query



Thanks

Translating Videos to Natural Language



Does not consider temporal sequence of frames.

[Venugopalan et. al. NAACL'15]